한국경영학회 1999년도 동계학술 발표회 제출논문

혼합 정수계획법을 이용한 도산 예측 문제에서의 분류 오차의 최소화

Minimizing the Misclassification Cost using Mixed Integer Programming in Bankruptcy Prediction Problem

조홍규, 한인구

서울 동대문구 청량리동 207-43 한국과학기술원 테크노 경영대학원, 우 130-012

Tel.: 02-958-3673, Fax: 02-958-3604, E-mail: hkjo@kgsm.kaist.ac.kr

1. 서론

기업의 도산은 매우 빈번하게 발생하며 경제적으로 여러 가지 문제를 발생시키는 현상이다. 특히, 국제통화기금 (IMF: International Monetary Fund)의 상황하에서 최근 국내에서 도산 기업의 수가 급격하게 증가하고 있는 실정이다. 그러므로, 기업의 도산에 대한 연구는 주의 깊은 관찰의 대상이 되어야 한다.

도산 예측은 경영학적 분류 문제의 한 분야이다. 경영학적 분류 문제는 기업 내에서 진행되는 의사 결정 과정 중 다양한 영역에 걸쳐 존재하고 있다. 도산예측, 회계방법 선정, 감사의견 결정, 대출 의사결정, 채권 및 신용 평가 등의 문제가 그 예이다.

분류란, 개체나 관찰치 들을 이미 선정해 놓은 집단으로 분류하는 것을 의미하는데, 전통적인 경영학 연구에서 중요한 연구분야로 자리잡고 있다. 분류의 궁극적 목적은 의미 있는 결과를 제공하고, 전문가의 판단을 대체하는 것이다. 개체나 관찰치를 구분하고 할당하는 방법을 분류 기법이라 부른다.

기존의 많은 연구들이 여러 가지 분류 기법을 적용하여 그 중에서 하나의 기법이 주어진 문제에 대해 좋은 성과를 보인다는 결과를 발표하였다. 그러나, 많은 연구 결과들이 특정한 문제에 대한 동일한 결론을 도출하지는 못하였고, 어떤 문제에 가장 적합한 기법이 무엇인지에 대해서는 아직 논란의 여지가 남아 있다. 그러므로, 여러 분류 기법의 비교를 통한 최적 기법에 대한 탐색 방법에 대한 대안으로 분류 기법들을 통합하는 방법론이 제시되고 있다.

본 연구는 상이한 분류 기법의 선형 통합 방법론을 제시하고 있다. 연구에서 제시된 방법론은 각각의 분류 기법을 통합할 수 있는 최적의 가중치를 찾아내고, 각 분류

기법의 가중 평균을 계산하여 최종 결과치를 산출하는 것이다. 제시된 방법론은 혼합 정수 계획법의 모양으로 구성되어 있다.

제안된 결합 방법론의 목적 함수는 총 분류 오차 비용 (total misclassification cost)을 최소화하는 것이다. 총 분류 오차 비용은 두 종류의 분류 오차의 가중 합계로 표현된다. 문제 해결 과정을 단순화하기 위하여 분류점 (cutoff point)을 고정하고 문제에 사용된 단계 함수 (threshold function)을 변환하였다.

혼합 정수 계획법의 형태로 구성된 문제는 분기 및 한계 기법 (branch and bound method)을 이용하여 해결하였다. 제시된 혼합 정수 계획법의 결과는 통합에 사용된 개별 기법의 결과보다 정확하게 사례를 분류하였다. 성과에 대한 비교는 통계적 방법을 이용하여 성과 차이의 유의도를 검증하였고, 제시된 방법론의 결과는 다른 개별 방법론의 결과보다 통계적으로 유의하게 좋은 성과를 나타내었다.

통합 방법론의 결과를 도출하기 위해 다양한 통계적 기법과 인공지능 기법을 이용한 도산예측 실험이 선행되었다. 사용된 통계적 방법론은 판별 분석, 로지스틱 회귀 분석이며 인공신경망 방법이 인공지능 기법으로 사용되었다.

2. 모형 설계

기존의 많은 연구들을 통해 다양한 종류의 요인들을 가진 문제에서 상이한 방법에 의해 예측된 결과의 가중 평균이 전체 모형의 성과를 증가시킨다는 결과가 발표되었다 (Mostaghimi, 1996). 가장 대표적인 시계열 예측 방법 중의 하나인 ARIMA 모형은 자기 회귀 모형 (AR: Auto regressive)과 이동 평균 모형 (MA: Moving average)의 선형 결합이다. 또한 분류 문제에 적용된 대부분의 수리 모형들은 독립 변수간의 가중 평균을 이용하고 있다. 본 연구는 도산 예측의 문제를 풀기 위한 각각의 분류 기법의 결과를 선형 결합하여 분류 오차를 최소화하는 방법론을 제시하였다.

각각의 분류 기법이 하나의 사례에 대해 출력값 O를 산출하였을 때, 결합된 출력값 C는 다음의 수식으로 표현된다.

$$C_i = \sum_{j=1}^{k} W_j \times O_{ij}$$

위 수식에서 i는 모형에 사용된 사례를 나타내며 1부터 n까지의 값을 가지고, j는 사용된 분류 기법의 순서를 표시한다. 도산 예측과 같은 이진 분류 문제에서는 각각의 사례가 두 개의 집단 중 어느 집단에 속하는지를 나타내기 위해 종속 변수를 0과 1의 두 개의 값으로 표시한다. 그러므로, 모형에 의해 출력된 결과값이 [0, *분류점*)의 범위에 있으면, 0의 집단으로 분류된 것이고, 결과값의 범위가 (*분류점*, 1]에 있으면, 1의 집단으로 분류된 것이다. 결과적으로 출력값 O_{ij}는 [0, 1]의 범위의 실수값이다.

가중치 W_j는 k개의 다른 분류 기법의 출력값을 결합하는 선형 결합식의 계수이다. 그러므로, O_{ij}의 가중 평균인 C_i도 [0, 1]의 범위의 실수값이 된다. 결합 출력값 C_i와 실제 사례의 집단값을 이용하여 사례가 정확하게 분류되었는지를 계산하고 그 결과를 총 분류 오차를 최소화하는 수리 계획 모형으로 구축한다.

분류 오차는 [표 1]에서 보여주는 바와 같이 두 가지 종류가 있다. 실제 사례의 종속 변수가 "1" 집단일 때, 예측 결과값이 "0" 집단인 경우와, 실제 "1"인 집단에 대해 "0"인 집단으로 예측하는 경우이다. 또한, 두 가지의 분류 오차는 상이한 위험 정도를 가지고 있다. 예를 들어, 금융 기관에서 고객 기업의 신용도를 판정하는 경우, 도산할 가능성이 큰 기업을 정상 기업으로 판단하여 대출했을 때의 위험 손실이, 정상 기업을 도산 가능성이 큰 기업으로 판단하여 대출을 거부했을 때의 손실보다 더 크다.

[표 1] 결합 예측치에 따른 분류 결과

IF	결합 예측치 C_i의 범위 = [0, *분류점*) ?		THEN	
	IF	실제 사례의 종속 변수 = 0 ?	THEN	*정확한 분류*
	ELSE	실제 사례의 종속 변수 = 1		*분류 오류*
ELSE	결합 예측치 C_i의 범위 = (*분류점*, 1]			
	IF	실제 사례의 종속 변수 = 0 ?	THEN	*분류 오류*
	ELSE	실제 사례의 종속 변수 = 1		*정확한 분류*
END IF				

총 분류 오차 비용을 MC라 정의하고, M_1을 종속 변수가 "1"인 사례에 대한 분류 오류의 수, M_0를 종속 변수가 "0"인 사례에 대한 분류 오류의 수라 정의하면 다음과 같은 관계가 성립된다.

$$MC = r_0 M_0 + r_1 M_1$$

위 관계에서 r_0은 종속 변수가 "0"인 집단의 분류 오차 위험이고, r_1은 종속 변수가 "1"인 집단의 분류 오차 위험이다. 분류 오류가 있었던 사례의 수를 나타내는 M_0과 M_1을 구하기 위해 단계 함수 (step function, threshold function) f를 도입하였다. 이 함수 f는 결합 예측치 C_i와 *분류점* 값의 차이를 0과 1의 값으로 변환시켜 준다. 최종적으로 수리 계획 모형의 목적 함수는 분류 오차 비용 MC를 최소화하는 것이다. 모형의 구조는 다음과 같다.

Minimize $\quad MC = r_0 M_0 + r_1 M_1$

s.t. $\quad C_i = \sum_{j=1}^{k} W_j \times O_{ij}$

$$M_1 = \sum_{i \in Group_0} f(Cutoff - C_i)$$

$$M_2 = \sum_{i \in Group_1} f(C_i - Cutoff)$$

$$f(x) = \begin{matrix} 0 & \text{if} & x \le 0 \\ 1 & \text{if} & x > 0 \end{matrix}$$

$$\sum_{j=1}^{k} W_j = 1$$

$$0 \le O_{ij}, C_i \le 1 \qquad \forall i, \forall j$$

$i=1,2,3,\cdots,n \quad j=1,2,3,\cdots,k$

위의 수리 모형에서 문제를 단순화하기 위해 두 가지의 가정을 추가하였다. 분류 오차의 상이한 위험 정도를 나타내는 r_0과 r_1이 같은 크기를 가지도록 하였다. 이 가정을 통해 r_0과 r_1을 목적 함수에서 제거할 수 있었다. 또한, 분류점 값을 0.5로 고정시키고, 단계 함수 f 대신에 오차 단위 Y_i를 사용하였다. 오차 단위 Y_i는 분류 오류의 수를 나타내는 M_0와 M_1의 합과 같다. 결국 최종적으로 다음과 같은 수리 모형을 구성하였다.

Minimize $\quad MC = \sum_{i=1}^{n} Y_i$

s.t. $\quad \sum_{j=1}^{k} W_j O_{ij} + 0.5 Y_i \geq 0.5$ 종속 변수가 "0"인 사례

$\sum_{j=1}^{k} W_j O_{ij} - \left(\sum_{j=1}^{k} O_{ij} - 0.5 \right) Y_i < 0.5$ 종속 변수가 "1"인 사례

$Y_i = 0 \text{ or } 1 \quad \forall i$

$\sum_{j=1}^{k} W_j = 1$

3. 문제 정의 및 실험

실험에 사용된 자료는 국내 중소기업 중 도산 기업과 정상 기업의 자료로 이루어져 있다. 자료에는 1993년부터 1995년까지 도산한 901개 기업이 포함되어 있고, 동수의 정상 기업 자료는 "신용보증기금"의 데이터베이스에서 선정되었다. 정상 기업의 자료는 업종에 따라 도산 기업과 같은 수의 사례가 포함되도록 Pair-matching 방법에 의해 선정되었다. 자료 중 도산 기업의 업종별 분포는, 중화학 공업 관련 기업이 436개, 경공업 관련 191개, 건설업 148개, 도매 및 소매업 64개, 기타 서비스업 62개로 이루어져 있다. 정상 기업의 자료 분포는 도산 기업의 분포와 같다.

실험의 독립 변수로는 재무비율이 사용되었다. 안정성, 수익성, 성장성, 현금 흐름, 활동성, 생산성의 6개의 범주에 걸쳐 총 67개의 재무 비율을 대상으로 독립 변수 선정 작업을 통해 10개의 재무 비율을 독립 변수로 사용하였다. 재무 비율 이외의 변수로 업종 분류, 업력, 기업의 종류 (상장, 외감, 등록, 일반) 등의 세 개의 변수가 추가 되었다. 독립 변수의 선정은 통계 모형의 단계적 선택법을 이용하여 이루어졌다. 독립변수의 목록은 [표 2]와 같다.

방법론의 일반성을 상호 확인하기 위해 전체 자료를 10개의 집단으로 다시 나누었다. 전체 샘플은 1622개의 학습용 샘플과 180개의 검증용 샘플로 이루어져 있다. 실험은 학습용과 검증용 샘플을 변화시켜 가면서 총 10회 반복 실험하였다. 다음 장에서 보여주는 결과는 10회 반복 실험의 평균값을 보여주고 있다.

2장에서 제시한 방법론을 실제 프로그램으로 구현하기 위해 "CPlex Mixed

Integer Programming" UNIX version 소프트웨어를 사용하였다. 구현된 프로그램은 "분기 및 한계 기법"을 이용하여 문제를 풀었고, 최대 분기 및 한계 노드 수는 20000으로 한정하였다.

[표 2] 독립변수 목록

안정성	자기자본 비율
	당좌 비율
	단기차입금 대 총차입금
수익성	총자본 경상이익률
	순금융비용 대 매출액
	이자보상배율
활동성	매출채권 회전율
성장성	총자산 증가율
현금 흐름	이자 등 지급 후 현금 흐름 대 부채 비율
생산성	부가가치율
비재무 자료	업종
	업력
	기업 종류

4. 실험 결과

실험의 결과는 각 기법별 성과 비교를 통해 분석되었다. 모형의 성과를 측정하는 방법은 다양하지만, 본 연구에서는 분류 문제의 성과 측정 수단으로 가장 많이 사용되고 있는 적중률을 사용하였다.[표 3]은 학습용 샘플과 검증용 샘플에서 개별 분류 기법들의 적중률과 본 연구에서 제시한 방법론의 적중률을 비교한 것이다.

[표 3] 각 기법 및 통합 방법론의 예측 적중률

	학습용 샘플				검증용 샘플			
	판별분석	로지스틱 회귀분석	인공 신경망	통합 방법론	판별분석	로지스틱 회귀분석	인공 신경망	통합 방법론
Fold 1	75.51	77.54	77.48	78.35	75.56	76.67	77.22	77.78
Fold 2	74.65	77.17	76.43	77.30	77.78	80.00	80.00	80.00
Fold 3	75.08	76.43	76.99	77.30	76.11	75.56	76.11	76.11
Fold 4	75.45	77.79	77.48	79.03	70.00	73.33	74.44	75.00
Fold 5	75.51	77.61	77.85	78.35	75.56	74.44	77.78	78.89
Fold 6	75.63	77.61	77.05	77.73	72.78	72.78	72.78	72.78
Fold 7	75.26	76.87	76.37	77.11	78.89	79.44	77.78	79.44
Fold 8	75.20	76.87	76.50	77.17	78.33	81.11	80.00	80.56
Fold 9	75.26	76.99	77.61	78.10	73.33	73.89	75.56	76.11
Fold 10	76.00	77.11	74.65	77.48	71.11	73.89	75.56	75.56
평균	75.35	77.20	76.84	77.79	74.94	76.11	76.72	77.22

[표 3]에서 보는 바와 같이 통합 방법론의 성과는 학습용 샘플과 검증용 샘플 모두에서 개별 분류 기법 보다 좋은 성과를 보여 주었다. 제시한 통합 방법론이 개별

기법의 예측 결과를 가중 평균하여 그 결과의 분류 오차를 최소화하는 것을 목적 함수로 하고 있으므로, 학습용 샘플에서 높은 성과를 얻으리라는 것은 설계 과정에서 이미 결과를 가정할 수 있었다. 한편 검증용 샘플에서의 성과는 통계적 방법을 이용하여 유의도를 검증하였다. 10개의 상이한 검증용 샘플의 성과를 이용하여 Paired t-test를 수행한 결과 통합 방법론의 예측력이 판별분석에 비해서는 1% 유의 수준에서, 로지스틱 회귀분석과 인공 신경망에 대해서는 5% 유의 수준에서 좋은 성과를 보여주었다.

5. 결론 및 제언

본 연구의 공헌은 상이한 기법의 결합을 통한 새로운 선형 통합 방법론을 제시한 것이다. 분류 오차 비용에 근거하여 방법론을 정립하였고, 실제 문제에 적용하기 위하여 몇 가지 가정을 함으로써 현행 수리계획법의 문제 해결 방법을 동원하여 문제를 해결할 수 있도록 만들었다.

본 연구의 한계점은 이원 분류를 위해 개발된 방법론을 다원 분류의 문제로 확장 시키지 못한 것이다. 도산예측의 문제는 종속변수의 값이 두 집단으로 분류되는 문제이고, 신용 평가, 어음 평가 등의 문제는 종속 변수의 값이 세 개 이상이 문제이다. 본 연구에서 제시된 방법론을 다원 분류 분제로 확장 시키기 위해서는 분류 오차 비용의 정의와 계산 과정의 변화가 요구되는 바이다.

또한 도산예측에 근거하여 실제 금융 기관에서 이루어지는 여신 결정 절차를 분석하는 연구가 향후 기대되는 바이다. 기업의 건전 정도를 도산예측을 이용하여 평가하고, 향후 대출 금액의 결정, 대출 이자율의 결정, 대출 위험의 효율적 관리 등의 문제와 연결하여 연구할 수 있을 것이다.

6. 참고 문헌

1. **Altman, E., G. Marco, & F. Varetto. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of Banking and Finance, 18, 505-529.**
2. **Banks, W., & P. Abad. (1991). An efficient optimal solution algorithm for the classification problem. Decision Science, 22, 1008-1023.**
3. **Jo, H., I. Han, & H. Lee. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. Expert Systems with Applications: An International Journal, 13, 97-108.**
4. **Jo, H., & I. Han. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. Expert Systems with Applications: An International Journal, 11, 415-422.**
5. **Mostaghimi, M. (1996). Combining ranked mean value forecasts. European Journal of Operational Research, 94, 505-516.**
6. **Myung, I. J., S. Ramamoorti, & A. D. Bailey, Jr. (1996). Maximum entropy aggregation**

of expert predictions. Management Science, 42, 1420-1436.

7. Troutt, M. D. (1995). A Maximum decisional efficiency estimation principle. Management Science, 41, 76-82.