

# 러프집합이론과 사례기반추론을 결합한 기업신용평가 모형

노태협\*·유명환\*\*·한인구\*\*\*

## <목 차>

I. 서론	V. 실증 분석
II. 기업신용평가모형	5.1 실험 데이터
III. 연구기법	5.2 실험 모형 설계
3.1 러프집합이론	5.3 결과
3.2 사례기반추론 (CBR, Case-Based Reasoning)	VI. 결론
IV. 러프집합-사례기반추론결합모형	참고문헌
	Abstract

## I. 서론

경기변동의 심화에 따른 기업의 부실채권 처리 문제가 은행을 포함한 금융기관들의 존립 여부를 결정짓는 현안으로 등장하였다. 이러한 문제는 그 동안 쌓였던 여러 불합리한 금융관행과 금융당국의 무사 안일한 의사결정에도 책임이 있겠지만, 무엇보다 적절한 기업신용평가 기준과 그에 맞는 합리적인 규모의 대출을 승인하는 시스템을 미리 구축해 놓지 못한 금융기관 스스로에게 일차적 책임이 있다고 하겠다.

기업의 신용에 대한 정확한 정보 요구가 은행과 몇몇 금융기관에만 국한되는 것은 아니다. 기업의 신용을 객관적이고 능동적으로 평가하여 그 결과를 해당 투자자들에게 제공하는 것은 금융시장의 건전한 발전을 위하여 필수적인 장치이다. 따라서 각 나라마다 혹은 국제적으로 활동하는 기업신용평가 전문기관들이 존재하기 마련이다. 기업신용평가 전문회사의 신용평가결과는 해당기업의 유가증권발행 또는 금융기관의 여신결정에서 대출여부 및 금리 등의 조건을 결정하는데 중요한 근거가 된다. 더욱이 금융시장이 개방되면서 국제금융시장에서 금융자본을 조달할 수 있는 시대가 도래했다. 따라서, 그 어느 때보다 기업의 신용을 합리적이고 객관적으로 평가하여

\* 한국과학기술원 테크노경영연구소 책임연구원, rohth@kgs.m.kaist.ac.kr

\*\* 한국산업은행, toto318@kgs.m.kaist.ac.kr

\*\*\* 한국과학기술원 테크노경영대학원 교수, ighan@kaist.ac.kr

그 결과에 따라 대출여부 및 한도, 그리고 각종 재무 투자안을 평가할 수 있는 과학적인 기업신용평가 시스템 구축이 절실한 시점이다.

기업의 도산예측 및 신용평가 모형 개발을 위해 다변량 판별분석(MDA), 회귀분석(Regression), 로짓(Logit), 프로빗(Probit) 등의 통계적 방법들을 주로 활용하여 왔다(Altman, 1968; Ohlson, 1980; Altman, 1983). 그러나, 1980년대 후반부터 귀납적 학습방법(Inductive Learning), 인공신경망(Artificial Neural Network), 유전자 알고리즘(Genetic Algorithm)과 같은 인공지능(Artificial Intelligence) 기법들의 우수성이 발표되면서 기업도산 및 채권평가 등에 이러한 방법론을 활용하기 시작하였다.

본 연구에서 대상으로 하는 기업신용평가는 기업도산예측, 대출승인결정과 같은 이진 분류(Binary Classification) 문제와 달리 다분류(Multi Classification) 문제이며, 그만큼 고도로 숙련된 신용평가 전문가들의 지식이 요구된다. 현재까지 통계와 인공지능 분야에서 이러한 기업신용평가 문제를 효과적으로 해결하기 위한 많은 연구가 진행되었다. 이러한 연구는 대개 정량적 분석을 기반으로 하고 있기 때문에 초기에는 회귀분석, 다변량 판별분석, 주성분 분석, 로짓 분석과 같은 통계적 방법들이 널리 이용되었으나 최근에는 인공신경망, 사례기반추론, 유전자 알고리즘 등과 같은 비통계적 방법들이 활발하게 사용되고 있다. 최근의 많은 연구들이 통계적 방법론이나 ID3, CART, C4.5같은 귀납적방법론(Inductive Learning)보다 인공지능기법에 의한 기업신용평가의 예측 정확도를 더 향상시킬 수 있음을 보여주고 있다. (Jo, Han and Lee, 1997; Kim and Han, 2000).

인공신경망을 포함한 여러 평가 모형의 활용 이전에 선행되어야 하는 작업으로 의미 있는 입력변수의 선택이 있다. 이는 해당 모형의 예측력 극대화과 설명력 향상을 위해 반드시 필요한 과정이다. 지금까지 연구되어온 변수 선정 기법으로는 판별 분석, 로짓 분석, 주성분 분석, 요인 분석, 그리고 러프집합 분석 등이 있다.

본 연구에서는 이러한 최근의 연구동향에 착안하여 기존에 여러 분야에서 응용되던 변수선정 기법과 예측 분류 기법의 두 가지 방법론을 결합한 새로운 형태의 기업신용평가 모형을 제안하고자 한다. 이는 집합이론에 그 근원을 두고 있는 러프집합이론(Rough Set Theory)과 기구축된 사례베이스를 통해 새로운 사례의 결과를 예측하는 사례기반추론(CBR, Case-Based Reasoning) 결합 모형이다.

비교적 최근에 연구가 시작된 러프집합이론을 기업신용평가 모형에 도입하게 된 이유는 무엇보다 러프집합이론에서 보여지는 집합과 그 집합에 대응되는 원소 사이의 독특한 소속개념 때문이다. 러프집합은 하위근사와 상위근사라는 모호한 경계영역(boundary region) 때문에 그 동안 많은 사람들이 퍼지집합(Fuzzy Set)과 혼동하여 왔다. 하지만 이 두 집합이론은 근본적으로 성격이 다르고 결과적으로 그 차이 때문에 러프집합이론을 본 연구에 사용하게 된 것이다.

고전적 집합이론을 집합의 경계 및 원소 소속의 이산성/연속성 여부에 의해 분류할 때 크게 보통집합(Crisp Set)과 퍼지집합(Fuzzy Set)으로 나눌 수 있는데, 러프집합은 근본적으로 보통집합에 그 기원을 두고 있다. 예를 들면, '키가 큰 사람들의 모임'이라는 퍼지집합은 집합 자체의 경

계는 모호하지만 172Cm, 180Cm와 같은 구체적으로 명시된 원소를 가질 수 있다. 물론 위 원소들이 '키가 큰 사람들의 모임' 집합에 포함되는 정도는 소속함수(membership function)에 의해 결정된다. 이런 면에서 러프집합은 퍼지집합과 정반대의 성격을 띠고 있다. 즉, 집합 자체는 그 범위가 보통집합처럼 뚜렷하지만 집합에 소속되는 원소들에 불확실성이 내포되어 있기 때문에 확실하게 소속 집합을 결정할 수 없는 것이다. 따라서, 명확하게 여러 등급으로 나누어지는 신용등급이라는 집합에 비해 정확하게 어느 등급에 속하는지 알 수 없는 기업이라는 원소들이 혼재되어 있는 현실 적용에서 이러한 러프집합이론이 적용 가능하다. 본 연구에서는 러프집합이론을 통해 선택된 주요 변수를 기반으로, 기업의 신용등급을 예측하기 위해 사례기반추론을 분류모형(Classifier)으로 사용한다. 상대적으로 적은 규모의 자료를 가지고 다등급체제로 이루어진 신용도의 다분류 문제에 높은 적중률을 보일 수 있으며, 과거 자료들이 가지는 주요 특성을 신규 데이터에 대해서도 효과적으로 반영할 수 있는 점을 고려하여 적용한다(Riesbeck and Schank, 1989; Slade, 1991; Kolodner, 1993).

본 논문은 다음과 같이 구성되어 있다. 2장에서는 기존 논문에서 기업신용평가 모형 개발을 위해 사용한 주요 기법들에 대해 알아보고, 3장에서는 이 연구에서 사용되는 러프집합이론과 사례기반추론의 이론적 배경을 제시한다. 4장에서는 러프집합모형을 통한 변수 선정과 사례기반추론의 예측모형 결합의 제안모형을 설명한다. 5장에서는 본 논문에서 제안하는 러프집합-사례기반추론(RS-CBR) 모형의 실효성을 검증하기 위한 다양한 실험 과정과 사용 데이터에 대한 설명 및 그 의의에 대해서 기술한다. 제안모형의 기업신용평가 분야에 있어서의 적용 타당성은 인공신경망 기반의 모형과 통계적 모형 등의 대조 실험군을 통해서 알아보게 될 것이다. 제안모형을 포함한 실험 모형에 대한 결과를 제시하고, 결과에 대한 원인을 비교, 분석이 포함된다. 6장에서는 전체 결과 및 시사점을 알아보고, 향후 기업신용평가에서 제안모형의 활용 방안과 연구의 한계점을 논의한다.

## II. 기업신용평가모형

기업의 도산예측이나 기업어음, 채권, 자산유동화 증권(ABS, Asset Backed Securities) 등의 신용등급 결정을 하기 위한 방법으로 전통적으로 통계적인 방법론이 많이 사용되어 왔다. 단순하게는 단일변량을 이용한 모델에서 다변량 판별분석(MDA), 주성분 분석(PCA), 로짓 분석(Logistic Regression)에 이르기까지 다양한 방법들이 존재한다.

단일변량 분석이란 단일변수 즉, 한 개의 재무변수를 이용해서 기업의 부실화와 미래의 재무 상황을 예측하고자 하는 방법이다. 이 모형은 정상기업과 부실기업의 재무비율 사이에 체계적인 차이가 존재하고 그러한 차이가 기업의 부도예측을 위해 구체적으로 이용될 수 있다는 두 가지 가정을 기초로 출발하고 있다. 이러한 단일변량을 이용한 통계적 방법론의 대표적인 연구는 Beaver(1966)의 예측 모형이다. Beaver는 재무비율을 이용한 단일변량 분석을 기업부실예측에

적용함으로써 부실예측분야에 통계적인 방법론을 적용한 최초의 연구자라고 할 수 있으며, 그의 연구는 1954년부터 1964년 사이의 79개의 부실기업과 79개의 정상기업을 표본으로 선정하고 프로파일분석(Profile Analysis)과 이원분류검정 (Dichotomous Classification Test)을 이용하여 재무비율에 의한 기업 부실 예측력을 평가하였다.

다변량 판별분석의 최초 연구는 Altman (1968)의 Z-Score 모형이다. Altman은 다섯 가지의 재무비율을 이용한 다변량 판별분석(MDA: Multivariate Discriminant Analysis) 모형을 개발하였는데, 이는 개별적으로 관찰되던 재무비율을 수리적으로 종합하고 단순화하여 기업부실예측의 전형적인 모형으로 인용되고 있다. 다변량 판별분석 이외에도 프로빗분석(Probit analysis), 다중회귀분석(Multiple regression analysis)과 같은 다양한 통계방법론이 예측정확도를 높이기 위한 방법론으로 많이 사용되어 오고 있다(Martin, 1977; Johnson, 1979; Ohlson, 1980; Hamer, 1983; Takahashi et al., 1984; Zmijewski, 1984; Gentry et al., 1985; Lee and Oh, 1990).

1980년대 후반부터 인공신경망 기법과 귀납적추론방법 등의 인공지능 기법들이 기업신용평가 모형에 적용되어 왔으며(Elmer and Borowski, 1988; Odom and Sharda, 1990; Cadden, 1991; Chung and Tam, 1992; Tam and Kiang, 1992; Miller et al., 1995; Kingdom and Feldman, 1995; Barbro et al., 1996; Jo et al., 1997), 최근에는 다양한 인공지능기법들의 특징들을 결합하여 통합 방법론을 통한 예측력 향상의 방법론들이 제시되고 있다(Lee and Kim, 1994; Yang et al., 1999; Shin et al., 2000; Mckee and Lensberg, 2002) (<표1> 참조).

<표 1> 기업신용평가모형 관련 문헌연구

통계적 신용평가기법		
1966	Beaver	단일변량분석
1968	Altman	MDA
1977	Martin	Logit Analysis
1979	Johnson	주성분분석
1980	Ohlson	Logit Analysis
1983	Hamerd	Logit Analysis
1984	Takahashi et al.	주성분분석, MDA
1984	Zmijewski	Probit Analysis
1985	Gentry et al.	MDA, Probit Analysis, Logit Analysis
1990	Lee and Oh	Bayesian
인공지능기법적용 신용평가기법		
1988	Elmer and Borowski	Expert system
1990	Odom and Sharda	Neural Network
1991	Cadden	Neural Network

1992	Chung and Tam	ID3, AQ, Neural Network
1992	Tam and Kiang	ID3, Neural Network
1995	Miller et al.	Neural Network, Fuzzy Theory
1995	Kingdom and Feldman	GA
1996	Barbro et al.	Neural Network, GA, Logit Analysis
1997	Jo et al.	Neural Network, CBR, MDA
<b>통합신용평가기법</b>		
1994	Lee and Kim	Hybrid Neural Network
1999	Yang et al.	MDA, Neural Network, Probabilistic Neural Network
2000	Shin et al.	CBR, Induction
2002	Mckee and Lensberg	GA, Rough Set Theory

### III. 연구기법

#### 3.1 러프집합이론

러프집합이론은 1982년에 Pawlak에 의해 처음 제시되었다(Pawlak, 1982; Pawlak & Slowinski, 1994; Pawlak, 1997). 이 이론에 쓰일 수 있는 부분은 속성들간의 상관성을 찾는 부분과 어떤 속성의 중요성을 파악하는 것, 그리고 일관성이 없는 데이터를 처리하는데 사용하는 것들이다. 이 접근 방법은 불확실성으로 가득한 데이터를 처리할 수 있다는 점에서 확률이론, 증거이론, 그리고 퍼지 이론과 공통점을 가진다. 이 이론은 세상에 모든 개체들은 그들이 가진 어떤 정보로써 집합을 지을 수 있다는 가정 하에서 시작된다. 동일한 정보의 외연을 가진 개체들은 그들의 정보로 인해 동일한 것으로 취급된다. 이러한 동질성 관계가 러프집합이론의 기초가 된다.

어떤 동질성을 가진 집합을 기본집합(Elementary Set)이라 하고 이들이 어떤 전체집합에 대한 지식의 기본 단위를 이룬다. 어떤 기본집합의 합집합이 되는 집합을 일반집합(Crisp Set)이라 하고 그렇지 않은 경우를 러프집합(Rough Set)이라 한다. 따라서 어떤 집합은 그 집합이나 보집합의 구성원으로 확실하게는 속하지 않을 수 있는 개체들 즉, 경계부분개체(Boundary-line Cases)를 가진다. 러프집합은 하위근사(Lower Approximation)와 상위근사(Upper Approximation)로 불리는 몇몇의 일반집합으로 표현될 수 있다. 전자는 확실성 있게 한 집합에 속하는 경우이고, 후자는 목표로 하는 집합에 속할 수도 있는 경우이다.

러프집합에서는 개체의 정보를 정보 테이블(Information Table)의 형태로 나타낸다. 행 부분은 개체들에 대한 부분이고, 열 부분은 관계 파악을 위한 속성들로 구성되어 있다. 그들이 교차하는

셀 부분에는 개체들의 속성값으로 채워져 있는 일반적인 데이터베이스의 테이블 형태를 띤다. 이를 수식의 형태로 나타내면 다음과 같다(Pawlak, 1997).

$$S = \langle U, Q, V, f \rangle$$

$S$ : information table(정보 테이블)

$U$ : a finite sets of universe(개체)

$Q$ : a finite sets of attributes(속성)

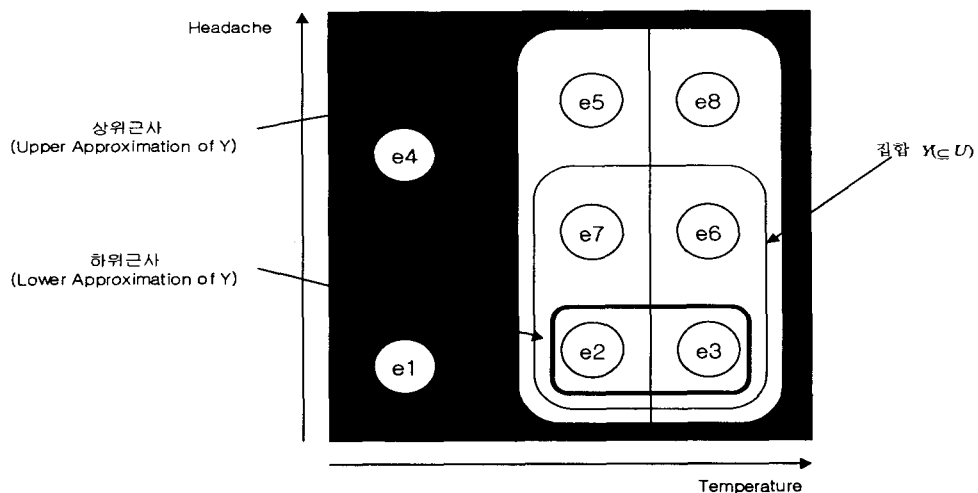
$V = \bigcup_{q \in Q} V_q, V_q$ : a domain of the attributes(속성값의 집합)

$f: U \times Q \rightarrow V$ : a total function such that  $\forall q \in Q, \forall x \in U f(x, q) \in V_q$  (Information function)

정보 테이블  $S = \langle U, Q, V, f \rangle$  에서 속성 집합인 부분집합  $P(\subseteq Q)$  을 잡고, 임의의 개체  $x, y(\in U)$  을 가지고 보았을 때,  $\forall q \in P, f(x, q) = f(y, q)$  의 조건을 만족하면 이 둘이 어떤 속성 집합  $P$  에 대하여 동질성 관계를 가진다고 말한다. 이로  $P(\subseteq Q)$  가 전체 개체집합에 대해서 쌍방의 동질성 관계를 형성시킬 때 이를 “ $P$ 에 대한 동질성 관계”라 하고  $IP$  로 표현한다.

전체 속성 집합  $Q$  중  $P(\subseteq Q)$  를 뺀고, 전체 개체 집합  $U$  중  $Y(\subseteq U)$  를 꺼내어 생각한다.  $P$ 에 대한 하위근사(Lower Approximation)와  $P$ 에 대한 상위근사(Upper Approximation)는 각각  $\underline{PY}$ 와  $\overline{PY}$ 로 일반적으로 표현하고 다음과 같이 나타낼 수 있다(Slowinski & Zopounidis, 1995) (<그림1> 참조).

$$\underline{PY} = \{x \in Y, I_p(x) \subseteq Y\}, \overline{PY} = \bigcup_{x \in Y} I_p(x)$$



<그림 1> 상위근사(Upper Approximation)와 하위근사(Lower Approximation)

또 다른 개념으로  $Y$  의  $P$ 에 대한 경계부분(P-Boundary of Set  $Y$ )이 있는데 이는 상위근사집단에서 하위근사집단에 속하지 못하는 확실성이 없는 부분을 가리키며,  $BN_P(Y)$  로 나타낸다. 그 의미는  $BN_P(Y) = \overline{PY} - \underline{PY}$  이다. 다시 말해서 하위근사집단  $\underline{PY}$ 는  $P$ 라는 속성집단에 의해 확실성 있게  $Y$  집합에 속할 수 있는 전체 개체 집합  $U$  의 모든 원소들의 집합이라 할 수 있고, 상위근사집합  $\overline{PY}$ 는 위와 같은 조건에서  $Y$  집합에 속할 수도 있는 집합이라 할 수 있다. 그 중에서 경계부분은 확실성을 가지고는  $Y$  집합에 속한다고 할 수 없는 집합이라 할 수 있다. 전체 개체 집합의 부분집합이 되는  $Y(\subseteq U)$  에서 어떤 속성집합  $P$  로써 ‘분류의 정확성’라는 것을 정의할 수 있고, 식은  $\alpha_P(Y) = \frac{Cardinality(\underline{PY})}{Cardinality(\overline{PY})}$  이다.

특정한 정보테이블  $S$  와 전체 속성 집합  $Q$  의 어떤 부분 집합  $P(\subseteq Q)$  를 고려하자. 그리고 전체 집합을 특정하게 분할한 집합  $Y = \{Y_1, Y_2, \dots, Y_n\}$  을 떠올린다. 이 분할은 속성 집합  $P$  에 속한 속성들과는 독립적인 것이다. 부분집합  $Y_i(i=1, \dots, n)$  은 분류  $Y$  의 집합들이 된다. 정보테이블  $S$  에서  $Y$  의 하위 그리고 상위 근사값은 다음과 같이 나타낼 수 있다.

$$\text{하위 근사: } \underline{PY} = \{\underline{PY}_1, \underline{PY}_2, \dots, \underline{PY}_n\}$$

$$\text{상위 근사: } \overline{PY} = \{\overline{PY}_1, \overline{PY}_2, \dots, \overline{PY}_n\}$$

이들을 이용해서 속성 집합  $P$ 에 의해 정의되는 분할  $Y$ 에 대한 “근사의 확실도”를 다음과 같이 나타낼 수 있는데, 이것은 러프집합 분석에서 분석의 기준이 되는 가장 중요한 값이다. 이 값은 정보 테이블 내에서 속성 집합  $P$ 가 어느 수준까지 의사결정 속성 클래스를 잘 구분할 수 있는가에 대한 지표로 그 수식은 다음과 같으며 정보테이블  $S$ 에서 속성 또는 속성집단 간의 종속적인 관계를 찾아내는 과정이 중요한 부분이다.

$$\gamma_P(Y) = \frac{\sum_{i=1}^n Cardinality(\underline{PY}_i)}{Cardinality(U)}$$

다른 중요한 문제는 속성군들 중 의미 있는 속성군을 선택하는 작업(Reduction)으로 이를 위해서 위의 “근사의 확실도”를 이용한다. 즉, 줄어든 변수군  $P(\subseteq Q)$ 가 전체 속성군  $Q$ 가 가질 수 있는 분류와 같은 정도의 “근사의 확실도”를 제공할 수 있도록 하는 방향으로 속성군의 원소 개수를 줄여간다. 여기에서 같은 정도의 “근사의 확실도”를 보장하면서 ( $\lambda_P(Y) = \lambda_R(Y)$ ), 가장 적은 속성의 수를 가진 속성군  $R(\subseteq P \subseteq Q)$  을 “속성군  $P$ 의  $Y$  -reduct”라 하고 줄여서

Reduct라 하는데, " $RED_Y(P)$ "라고 표기하도록 한다. 일반적으로 정보테이블은 하나 이상의 Reduct를 가진다. 모든 Reduct의 교집합이 되는 속성들의 집합을 "속성군 P의 Y-core"라 부르며, " $CORE_Y(P)$ "라고 표기하는데 식으로는 " $CORE_Y(P) = \bigcap RED_Y(P)$ "로 표현할 수 있다.

기존의 연구에서 러프집합이론은 주로 신용등급 결정을 포함한 기업평가에 쓰여 왔다. 가장 먼저 발표된 것은 39개의 기업 샘플을 이용해서 신용평가를 위한 간단한 규칙을 얻어낸 것이다 (Slowinski & Zopounidis, 1995). 그 후에 그리스의 기업 인수 가능성을 러프집합이론을 이용하여 평가한 후 이를 다변량 판별분석의 성과와 비교한 연구가 있다(Slowinski, Zopounidis, & Dimitras, 1997). 또한 기업 부실 예측을 위해 러프집합 분석을 사용한 후 그 성과를 로짓 모형의 성과와 비교한 논문이 있다(Dimitras, Slowinski, Susmaga, & Zopounidis, 1999).

이외에 다른 모형과의 결합을 위해 러프집합이론을 응용한 사례도 있었는데, Hashemi et al. (1998) 은 신경망의 전처리 과정으로 러프집합 분석을 사용했고 이를 위해 2-Dimensional Reduction을 통해 최적 샘플과 속성으로 정보 테이블을 구성한 후 실험하는 방법을 택하였다. 궁극적으로 이 모형을 이용해서 은행의 지배구조를 예측하려고 하였다. 이와 유사한 방식을 이용하여 김창연(1999)은 기업 부실 예측에 러프집합 분석을 적용하였다. 또한 인공신경망과 다변량 판별분석에서 상이하게 나타난 결과들에 대해 러프집합 분석을 이용하여 그 결과들을 통합하는 방법론도 제안된 바가 있다(박기남, 이훈영, & 박상국, 1999).

### 3.2 사례기반추론(CBR, Case-Based Reasoning)

사례기반추론은 새로운 요구에 대응하는 과거의 해답을 채택하거나, 과거의 사례를 이용하여 새로운 상황을 설명하거나, 과거의 사례로 새 해답을 평가하거나, 또는 새로운 상황을 이해하거나 새로운 문제에 대한 적당한 해답을 찾기 위해 선례로부터 추정하는 것을 의미한다(Kolodner, 1993). 사례기반추론의 기본원리는 전문가가 복잡한 문제를 풀기 위해서 과거에 문제를 해결하였던 경험을 바탕으로 비슷한 사례를 추출하여 주어진 문제를 풀고자 하는 유사추론, 혹은 경험적 추론을 하는 것과 같은 것이다. 그러나, 인간의 기억용량의 한계로 인하여 사례의 수가 증가하면 사람은 현재 주어진 문제와 가장 비슷한 사례를 찾지 못하고 최근에 해결했던 사례나 가장 기억에 남는 사례를 찾아서 문제를 해결하고자 한다. 그러므로 사례기반추론 시스템은 이러한 인간두뇌의 한계를 극복하도록 하여 전문가의 문제해결 기술을 증진시킨 시스템이라고 할 수 있다.

사례기반추론에 관한 연구는 매우 활발히 일어나고 있으며 최근 몇 년간 문제해결 학습의 도구로 많이 사용되고 있다. Brown과 Gupta(1994)에 따르면 최근에는 예산작성, 비용추정, 자산평가, 포트폴리오 관리, 기업의 채권등급 결정, 위험평가, 내부통제평가, 부정적발 등의 재무 및 회계분야에서 활발하게 적용되고 있다.

사례기반추론을 실행하기 위해서는 먼저 사례에 대한 표현방법과 사례에 대한 색인방법을 선택하여 사례기반추론의 구조를 결정하여야 한다. 사례의 표현방법을 결정한다는 것은 주어진 영역의 문제를 풀기 위해서 상황의 어떤 단면을 선택하고, 또 어떤 식으로 이들을 표현할 것인가를

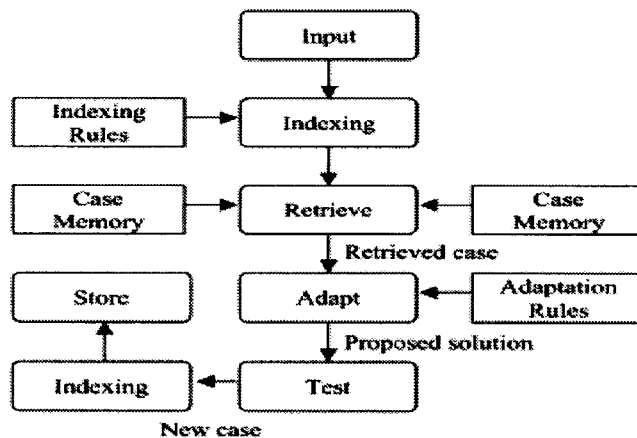


결정하는 것이다. 사례를 표현하는 방법에는 사례의 특성을 정의하는 속성(attribute)의 단순한 집합으로 사례를 표현하는 특징표현(feature representation) 방법, 속성의 값을 포함하는 슬롯(slot)의 집합을 구성된 프레임(frame)이 하나의 노드(node)가 되어 계층적 구조(hierarchical structure)로 사례를 표현하는 프레임 방법, 그리고 프레임의 확장된 개념으로 사례를 표현하는데 있어서 규칙, 프레임, 추상화(abstraction), 상속(inheritance) 등과 같은 개념을 혼합 적용하여 사례를 표현하는 방법 등이 있다. 사례에 색인을 붙이는 것은 특정한 사례가 주어졌을 때 사례베이스에서 유사한 사례를 효율적으로 찾기 위하여 사례베이스의 사례를 조직화하고 표식(label)을 붙이는 것을 의미한다.

사례기반추론에 관한 기존의 연구(Barletta, 1991; Kolodner, 1993; Brown & Gupta, 1994; Aamodt & Plaza, 1994)에서 제시한 사례기반추론의 과정은 각각 약간의 차이를 보이나 기본 골격은 일치한다. 먼저 새로운 문제가 주어졌을 때 사례베이스에서 일정한 유사성의 척도에 부합하는 과거의 사례를 추출한 후, 추출된 사례들을 미리 정해진 알고리즘에 적용하여 사례를 재사용(reuse)하고 해결에 이용한다. 만약 추출된 사례에 의한 해답이 새로운 문제 해결에 적합하지 않으면 이를 수정(revise)하고 새로운 해답을 제시하고 이를 다시 새로운 사례로 저장(retain)한다는 것이다. Brown과 Gupta(1994)는 사례에 색인을 붙이는 방법으로서 사례의 관련적 추출(associative retrieval)을 위해 사례의 특징에 개별적으로 색인을 붙이는 방법과 사례의 계층적 추출(hierarchical retrieval)을 위해 사례의 특징 중에서 사례의 공통된 성질을 상위 노드에 두고 하위 노드로 갈수록 서로 다른 성질을 가지도록 배열하여 계층적으로 색인을 붙이는 방법을 제시하고 있다.

위의 과정을 다음의 세 단계로 요약하였다(<그림 2>참조).

- 1단계: 재생(remembering) - 사례 추출(case retrieval)
- 2단계: 적용(applying) - 사례의 이용(case usage)
- 3단계: 학습(learning) - 사례집합 개선(case-base enhancement)



<그림 2 > 사례기반추론의 추론 과정

## IV. 러프집합-사례기반추론결합모형

기업의 신용평가 모형 개발에서 가장 어려운 작업 중의 하나는 무수한 신용지표들 중에서 해당 모형에 적합한, 즉 정확하게 사례 기업들의 특징을 알려줄 수 있는 변수를 선택하는 것이다. 그러한 변수들을 선택하는 데 있어서 전통적으로 사용되어 온 방법들로는 단변량 t-test, 1-way ANOVA, 판별분석 및 로짓분석의 stepwise method 등을 들 수 있다. 하지만 이러한 통계적 방법론에는 반드시 지켜져야 하는 엄격한 가정들이 뒤따르고는 했다. 따라서 최근에 이르러 이러한 통계적 가정을 완화하거나 아예 무시할 수 있는 방법론들이 활발히 연구되고 있고, 본 연구에서 사용하는 러프집합이론 역시 그 중의 하나로 볼 수 있다.

본 논문에서 기업의 신용등급을 예측하기 위한 분류모형(classifier)으로 사용한 것은, 상대적으로 적은 규모의 자료를 가지고도 높은 적중률을 보이면서 과거의 자료들이 가지는 주요 특성을 신규 데이터에 대해서도 효과적으로 반영하는 것으로 알려진 사례기반추론(CBR, Case-based Reasoning)이다. 또한 실제 러프집합 분석의 변수 선정 능력이 어느 정도인지를 알아보기 위하여 판별 분석이나 주성분 분석과 같은 전통적인 통계기법뿐 아니라 인공지능망 같은 대표적인 인공지능 기법들과의 적중률을 비교하였다.

최적의 변수를 선정하는 것(optimal feature selection)은 뛰어난 예측 능력을 갖춘 분류 모형(classifier)을 개발하는 데 있어 가장 중요한 과정의 하나이다. 최적의 변수 선정이란, 예측 모형(predictor)에 사용되는 최초의 입력 변수군과 동일하거나 혹은 더욱 우수한 성능을 나타내면서 동시에 미리 정의된 변수 선정 기준을 만족시킬 수 있는 변수군을 추출하는 것을 의미한다(Swiniarski & Hargis, 2001). 이러한 최적 변수군은 일반적으로 예측 모형(predictor)의 성격에 따라 많이 달라진다.

현재까지 알려진 변수 선정 방법론은 크게 다음과 같은 두 가지 흐름의 하나로 귀속시킬 수 있다.

오픈-루프 방식(open-loop, filter, front-end methods)에서는 선정된 변수들이 구체적으로 전체 예측 모형에 대해 어떠한 영향을 미칠 것인지에 대해 고려하지 않는다. 즉 새로운 변수군의 선정 과정과 선정된 변수군을 이용하는 예측 모형을 두 개의 독립적인 모형으로 간주하므로 선정된 변수군으로 인해 예측 모형의 적중률에 어떠한 변화가 나타나는지 등의 구체적인 성과지표에는 관심을 갖지 않는다. 그 대신 종속변수의 군집도를 나타내는 분산이나 공분산을 최대화 하는 등의 자체 변수 선정 기준을 충족시키는 과정에서 최적의 변수를 찾게 된다. 따라서 이러한 범주에 속하는 변수 선정 방법들은 최종 예측 모형의 결과에 따른 어떠한 피드백도 받지 않게 된다.

클로즈드-루프 방식(closed-loop, wrapper, classifier feedback methods) 혹은 예측모형 피드백 방식은 말 그대로 입력 변수 선정의 기준 자체를 최종 예측모형의 적중률을 높이는 것으로 하고, 연속적인 향상을 위해 계속해서 그 결과를 피드백 받게 된다. 따라서 오픈-루프 방식에 비해 훨씬 직관적인 기준을 가지고 있고 현실 세계의 일반적인 요구 수준에 좀 더 부합한다고 볼

수 있다. 선택된 변수군은 끊임없이 예측모형의 적중률을 테스트하면서 최초의 변수군과 그 결과를 비교하면서 최적의 성능을 보이는 변수군을 선택하게 된다. 이러한 클로즈드-루프 방식은 예측모형의 반복되는 피드백에 의해 일반적으로 그 성능은 오픈-루프 방식에 비해 뛰어난 것으로 알려져 있지만, 예비 변수 선정에서 최종 예측 모형까지를 하나의 루프로 보고 그 결과에 따라 임의의 선택을 반복하여야 하므로 그 연산량과 계산에 소요되는 시간이 큰 부담으로 작용한다.

본 논문에서 사례기반추론의 입력변수 선정을 위해 제안한 러프집합이론은 그 자체의 변수 선정 기준을 갖고 오직 그 기준에 맞추어 변수를 선택한 후 예측모형에 제공하는 대표적인 오픈-루프 방식의 변수 선정 방법이다. 러프집합이론을 기업신용평가 모형에 도입하게 된 이유는 무엇보다 러프집합이론에서 보여지는 집합과 그 집합에 대응되는 원소 사이의 독특한 소속개념 때문이다. 러프집합은 하위근사와 상위근사라는 모호한 경계영역(boundary region) 때문에 그 동안 많은 사람들이 퍼지집합(Fuzzy Set)과 혼동하여 왔다. 하지만 이 두 집합이론은 근본적으로 성격이 다르고 결과적으로 그 차이 때문에 러프집합이론을 본 연구에 사용하게 된 것이다. 명확하게 여러 등급으로 나누어지는 신용등급이라는 집합에 비해 정확하게 어느 등급에 속하는지 알 수 없는 기업이라는 원소들이 혼재되어 있는 현실 적용에서 이러한 러프집합이론이 적용 가능하다. 본 연구에서는 러프집합이론을 통해 선택된 주요 변수를 기반으로, 기업의 신용등급을 예측하기 위해 사례기반추론을 분류모형(Classifier)으로 사용한다. 상대적으로 적은 규모의 자료를 가지고 다등급체계로 이루어진 신용도의 다분류 문제에 높은 적중률을 보일 수 있으며, 과거 자료들이 가지는 주요 특성을 신규데이터에 대해서도 효과적으로 반영할 수 있는 점을 고려하여 러프집합을 이용한 변수 선정을 기반으로 사례기반추론 방법론을 결합하여 모형을 구축한다 (Riesbeck and Schank, 1989; Slade, 1991; Kolodner, 1993).

## V. 실증 분석

### 5.1 실험 데이터

본 논문에서는 보다 효과적인 기업신용평가 모형 개발을 위하여 러프집합이론과 사례기반추론을 결합한 모형을 제안하고자 한다. 이러한 결합모형의 성과를 증명하기 위하여 우리나라의 기업신용평가기관에서 수집한 1,295개의 기업신용 평가 자료를 기초로 실험을 하고 그 결과를 기존의 판별분석, 주성분분석 및 인공신경망 방법에 의한 결과와 비교하였다. 실험결과, 통계적으로도 유의하고 실무적인 관점에서도 의미가 있는 기업신용평가 결과를 유도할 수 있었다.

본 데이터는 기업신용평가 자료 중 제조업에 종사하며 KOSPI에 상장되어 있거나 KOSDAQ에 등록되어 있는 1295개의 기업을 표본데이터로 선정하였다. 기업의 신용등급은 신용정보기관의 공시 자료를 이용하였고, 실제 해당 기업들의 재무제표 자료는 한국상장회사협의회에서 제공

하는 상장협 데이터베이스에서 추출하였다. 표본을 굳이 제조업 소속 기업으로 한정하는 이유는 금융업과 서비스업 등 전체 업종을 망라하는 표본에서는 보다 정확한 모형을 구축하기 힘들뿐더러 각 업종별 특성이 많은 변수에서 상이하게 나타나므로 논리적으로 설명하기 힘든 결과를 초래할 수 있을 것으로 보았기 때문이다. 이들 기업의 신용등급은 크게 A1, A2, A3, B, C의 5등급으로 나뉘고 있는데, 본 연구에서는 편의상 A1은 1, A2는 2, A3는 3, B와 C는 4로 표기하였다. B와 C 등급을 하나의 등급으로 묶은 이유는 표본의 분포에서도 나타나듯 C등급을 받는 기업이 실제로 거의 없기 때문이다. C등급을 받은 기업들은 이미 도산한 기업이거나 향후 도산이 거의 확실시 되는 기업들로, 재무구조가 좋지 않은 대부분의 기업에 대해서 그 하한으로 B등급을 부여하고 있다. 따라서 2002년 기업평가에서 나타나는 극소수의 C등급 기업을 B등급과 함께 4등급으로 설정하여도 실제 평가와 큰 차이는 없을 것으로 예상할 수 있다. 표본 중 변수 값이 누락된 기업은 제외하였으며, 모형의 구축을 위한 학습용 데이터(training data)로 각 등급의 80%에 해당하는 데이터를 그리고 나머지 20%를 검증용 데이터(validation data)로 사용하였다. 모든 기업이 적어도 한 번은 모형의 학습과 검증을 위해 사용될 수 있도록 5-fold cross validation을 실시하였고, 해당 실험(experiment)의 결과와 함께 마지막에는 5개의 실험을 종합한 결과를 제시하였다. 최종적인 표본 기업의 신용평가 등급 분포는 <표 2>과 같다.

<표 2> 실험에 사용된 표본 기업의 신용평가통계 (제조업, 2002년)

등급	Training	Validation	합 계	백분율
1	82	24	106	8.2%
2	442	110	552	42.6%
3	266	66	332	25.6%
4	245	60	305	23.6%
합 계	1035	260	1295	100%

주) 1등급 → A1, 2등급 → A2, 3등급 → A3, 4등급 → B & C

기업신용평가 요인은 재무제표로부터 추출되는 재무변수와 기업의 형태, 산업 및 환경요인 등의 비재무변수로 나누어 볼 수 있다. 본 논문에서는 비록 계량적인 재무변수를 중심으로 하여 분석을 하였으나, 비재무변수가 가지는 추가적인 정보특성을 반영하기 위하여 측정 가능한 비재무변수를 추가하여 분석한다. 분석대상이 되는 재무변수는 기존 연구에서 기업의 신용등급을 결정하는데 유의한 것으로 밝혀진 39개의 변수(규모지표 10개, 수익성 지표 12개, 안정성 지표 10개, 현금흐름 지표 4개, 생산성 지표 3개, <표3> 참조)를 1차적으로 고려하였고(이건창, 한인구, & 김명중, 1996; Shin & Han, 1999; 박기남, 이훈영, & 박상국, 2000; Huang & Tseng, 2004), 각각의 방법론에 따라 고유의 변수 선정 과정을 거쳐 그 개수를 줄였다.

실험에 사용된 자료의 기초 통계분석 및 판별분석, 주성분분석을 위해 SPSS 11.0을 사용하였고, 인공신경망은 NeuroShell 2.0, 사례기반추론은 KATE 5.02, 그리고 리프집합 분석은 ROSETTA 1.4.41을 사용하였다.

<표 3> 실험에 사용된 변수의 기초통계량

변수 번호	변수명	평균	표준편차	최소값	최대값	단위
X1	총자산	326760552.43	974301004.16	4871587	7254476523.36	원
X2	유형자산	138031856.24	467951010.78	62905	3540379377.79	원
X3	고정자산	218509139.74	752048463.10	592178	5955596210.17	원
X4	자기자본	150030169.47	467468517.37	19185	4242837370.43	원
X5	매출액	306981010.40	985699039.37	1279735	6869244627.83	원
X6	부가가치	67421.62	228544.51	375513.95	2000146.52	원
X7	총부채	171673470.59	513319735.23	115819	3313858931.58	원
X8	감가상각비	2645342.12	17721460.02	494	267935486.49	원
X9	1인당매출액	427.30	436.30	10.3	2882.36	원
X10	영업이익	20377464.88	83410690.36	705615897.6	765361921.77	원
X11	당기순이익	10599242.00	73479618.85	704567122.7	741884427.25	원
X12	주당순이익	1084.99	3350.57	13912	17085.33	원
X13	업력	23.50	14.57	1	68.05	년
X14	총자본사업이익률	4.29	10.78	30.79658731	39.03	비율
X15	유보액대비율	38.00	29.38	75.06799026	116.42	비율
X16	금융비용부담률	1.22	4.18	25.23090075	28.14	비율
X17	금융비용대부채비율	0.00	0.00	0.001994538	0.00	비율
X18	금융비용대총비용비율	0.94	2.84	9.718768285	11.59	비율
X19	감가상각비대총비용비율	4.47	3.79	0.04	18.42	비율
X20	고정자산구성비율	48.56	19.11	2.21	98.14	비율
X21	차입금의존도	27.18	17.72	0.01	83.97	비율
X22	자기자본구성비율	53.97	21.04	0.24	99.07	비율
X23	재고자산대유동자산비율	24.20	16.10	0.01	73.77	비율
X24	단기차입금대총차입금비율	58.08	24.09	0.1	99.94	비율
X25	현금흐름대부채비율	12.07	36.48	143.5957547	165.06	비율
X26	현금흐름대고정부채비율	0.62	12.87	269.8968657	110.20	비율
X27	현금흐름대총자본비율	3.35	10.41	32.31392371	36.75	비율
X28	총자본투자효율	16.57	29.95	234.4	463.30	비율
X29	고정부채대자본금비율	1.56	3.48	0.001397623	68.22	비율
X30	부채비율	164.46	387.23	0.94	4361.76	비율
X31	이자보상배율	2.80	111.75	1291.330069	1258.42	비율
X32	총자본순이익률	0.67	15.45	59.8833453	57.73	비율
X33	자기자본순이익률	5.13	51.50	390.8357134	381.67	비율
X34	자본금영업이익률	53.04	128.77	512.05	787.50	비율
X35	자본금경상이익률	32.50	156.93	716.8245249	791.91	비율
X36	매출액총이익률	19.15	15.14	31.59095783	69.72	비율
X37	매출액경상이익률	5.86	36.62	192.8980176	175.27	비율
X38	(자본+고정부채)/고정자산	1.83	1.75	0.176434817	33.60	비율
X39	(영업활동으로인한현금흐름 현금배당)/(고정자산+운전자본)	0.04	0.26	2.007845094	3.95	비율

## 5.2 실험 모형 설계

제안모형의 기업신용평가 분야에 있어서의 적용 타당성을 인공신경망 기반의 모형과 통계적 모형 등의 대조 실험군을 통해서 검증하였다. 러프집합-사례기반추론결합모형의 신용평가 모형에서의 예측력 향상에 대한 실증 분석을 위하여 우선 실험에 사용되는 방법론 각각의 단일 모형에 대한 실험을 수행하며, 결합모형간의 예측력 향상을 살펴보기 위하여 인공신경망 결합모형을 사례기반추론 결합모형과 비교하였다. 본 논문에서 기업의 신용등급을 예측하기 위한 분류모형(classifier)으로 사용한 것은, 상대적으로 적은 규모의 자료를 가지고도 높은 적중률을 보이면서 과거의 자료들이 가지는 주요 특성을 신규 데이터에 대해서도 효과적으로 반영하는 것으로 알려진 사례기반추론(CBR, Case-Based Reasoning)이다. 또한 실제 러프집합 분석의 변수 선정 능력이 어느 정도인지를 알아보기 위하여 판별 분석이나 주성분 분석과 같은 전통적인 통계기법뿐 아니라 인공신경망 같은 대표적인 인공지능 기법들과의 적중률을 비교하였다. <표 4>는 본 실증분석에서 수행하게 되는 실증 분석 모형들에 대하여 정리한 표이다.

<표 4> 실증분석 모형 요약

종 류	모 형	설 명
단일모형	MDA	· 판별분석모형
	RS	· 러프집합모형
	ANN	· 신경망모형
	CBR	· 사례기반추론모형
인공신경망 결합모형	MDA-ANN	· 판별분석변수선정-신경망예측모형
	PCA-ANN	· 주성분분석변수선정-신경망예측모형
	RS-ANN	· 러프집합변수선정-신경망예측모형
사례기반추론 결합모형	MDA-CBR	· 판별분석변수선정-사례기반추론예측모형
	PCA-CBR	· 주성분분석변수선정-사례기반추론예측모형
	RS-CBR	· 러프집합변수선정-사례기반추론예측모형

### 5.2.1 단일 모형

판별분석(MDA, Multivariate Discriminant Analysis) 모형은 판별분석을 실시하기 이전에 1차적으로 일원 분산분석(1-way ANOVA)과 상관분석(Correlation Analysis)을 통해 39개의 변수군 중에서 17개의 변수를 분석에서 제외하였다. 최종적으로 22개의 변수를 사용하여 판별분석을 실시하였다. 22개의 변수 중에서 판별 모형의 구축에 도움이 되는 변수를 선정하기 위하여 판별분석의 stepwise 옵션을 이용하였다. 그 결과 각 실험(experiment) 별로 12~15개의 변수가 판별모형의 독립변수로 선정되었다.

러프집합 분석은 인공신경망이나 사례기반추론 같은 모형을 위한 변수 선정의 기준으로도 사용될 수 있지만, 그 자체로도 하나의 훌륭한 분류 모형이 될 수 있다. 특히 러프집합 분석의 산출물로 나타나는 IF - THEN 형식의 규칙은 다양한 형태의 의사결정 지원시스템 구축에 응용될 수 있어 매우 유용하다. 본 논문에서는 자료의 범주화로 불리안 추론(Boolean Reasoning)을, 자료 및 변수의 축약(horizontal & vertical reducing)을 위해서는 유전자 알고리즘을 사용하였다. 위 조합은 실제 자료의 범주화와 축약 과정에서 다른 조합에 비해 가장 성과가 뛰어났음을 밝혀준다. 그리고 유전자 알고리즘으로 리덕트를 찾아낼 경우 일반적으로 가용한 모든 리덕트를 찾아내므로 학습 데이터(training data)에 대해서는 거의 100%에 가까운 적중률을 보여준다. 이는 유전자 알고리즘을 이용할 경우 반복되는 교배(crossover)와 돌연변이 (mutation)를 통해 전체 최적해(global optimum)에 가까운 해를 찾아주는 것과 유사하다고 할 수 있다.

기존 연구에서 발췌한 39개의 변수에 대해 전처리 과정을 거치지 않고 모든 변수를 입력변수로 사용하여 인공신경망 모형 및 사례기반추론 단일 모형을 생성하였다. 그 결과 비교적 높은 적중률을 보여주고 있는데, 이는 39개의 변수가 이미 기존 연구에서 기업의 신용등급을 결정하는데 유의한 것으로 판명된 것이어서 이러한 결과가 나타난 것으로 추정하고 있다.

### 5.2.2 인공신경망 결합 모형

판별 분석, 주성분 분석, 러프집합 분석을 통해 선정된 세 가지 입력 변수군의 적합도를 검증하기 위해 인공신경망을 분류 모형(Classifier)으로 사용하는 결합 모형을 생성하였다. 즉 MDA-ANN, PCA-ANN, 그리고 RS-ANN이라는 세 가지의 모형에 대해 그 적중률을 평가하는 것이다. 본 실험에 앞서 입력변수 축소에 따른 모형의 전반적인 성과 변화를 알기 위하여 변수를 거르지 않고 39개의 모든 변수를 입력 변수로 사용한 인공신경망 모형도 실험하였다.

실험에는 NeuroShell 2를 사용하였고 Test set의 Calibration interval을 200으로, MAE(Minimum Average Error)에 도달한 후에 30,000회의 이벤트를 수행하도록 설정하였다. 은닉층(hidden layer)의 노드(node) 개수는 0.5n, n, 1.5n, 2n(n은 입력층의 노드 개수) 중에서 가장 성과가 좋은 것을 선택하였다. 그리고, 사례기반추론 모형과의 형평성을 유지하기 위해 모든 변수들은 3개의 범주로 등분위 범주화(Equal frequency binning)하여 입력하였다.

판별 분석의 stepwise 옵션을 이용하여 선정된 판별 모형의 독립변수들을 인공신경망 모형의 입력변수로 사용하였다. 전반적으로 순수 판별 모형보다는 높은 적중률을 보이지만, 러프집합 분석보다는 낮은 성과를 보이므로 결합 모형(hybrid model)으로서는 그리 좋은 성과라고 할 수 없다. 본 논문에서는 총 변동의 약 80%까지를 설명할 수 있도록 주성분을 선택하였고, 그 결과 각 실험에서 12 ~ 13개의 주성분으로 축약할 수 있었다.

러프집합 분석의 목적은 그 자체의 최종 산출물인 IF - THEN 형식의 규칙 생성을 통한 신규 데이터의 예측뿐 아니라 함께 제공되는 리덕트(reduct)와 코어(core)라는 변수군을 통해 변수간 중요성의 순서 및 정도를 파악할 수 있는 것이라고 밝힌 바 있다. 이 중에서 코어(core)란, 모든 리덕트(reduct)에 포함되어 있는 가장 중요한 의미를 가지는 변수들의 집합으로서 본 연구의 모

티브가 된다.

본 모형에서는 러프집합 분석을 통해 생성된 코어와 코어에 근접하는 변수들을 인공지능망의 입력변수로 사용하였다. 코어가 아닌 코어에 ‘근접’하는 변수라고 밝힌 이유는 실제 완벽하게 코어가 되는 변수는 많지 않기 때문에 가장 리덕트에 많이 포함되는 변수들의 순으로 정렬하여 입력변수군을 형성하였기 때문이다. 입력변수의 개수는 최초 1개에서 다음 순위의 코어를 하나씩 추가할 때마다 나타나는 적중률 분포의 추이를 보면서 적중률이 급격하게 감소하거나 혹은 일정 수준 이상의 증감이 나타나지 않을 때까지로 하였다. 하지만 변수의 개수가 지나치게 많아지면 모형으로서의 의미가 없어지기 때문에 최대 20개를 넘지 않도록 하였다. 실험 결과, 본 모형의 적중률을 최대화 시켜주는 코어 집합은 러프집합-사례기반추론(RS-CBR) 모형에서 선정된 코어 집합과 동일한 것으로 나타났다. 따라서 실제 선정된 입력변수에 관한 내용은 러프집합-사례기반추론(RS-CBR) 모형의 결과를 참조하면 될 것이다.

러프집합-사례기반추론(RS-CBR) 모형과의 성과를 비교하기 위하여 공히 자료는 3등분위로 범주화(Equal frequency binning) 하였고, 유전자 알고리즘(GA, Genetic Algorithm)을 이용하여 리덕트와 규칙을 생성하였다.

### 5.2.3 사례기반추론 결합 모형

인공지능망 결합 모형의 실험과 동일한 조건 하에서, 인공지능망 대신 사례기반추론을 분류 모형(Classifier)으로 놓고 세 가지 변수 선정 기법의 유용성을 검증하였다. 본 실험에 앞서 입력 변수 축소에 따른 모형의 전반적인 성과 변화를 알기 위하여 변수를 거르지 않고 39개의 모든 변수를 입력 변수로 사용한 사례기반추론 모형도 실험하였다. 실험은 KATE 5.02를 사용하였고, 유사 사례 추출을 위하여 10-NN(Nearest Neighbor)를 적용하였다. 사례기반추론은 입력변수의 중요도에 따라 가중치 차이를 둬으로써 그 성능을 향상시킬 수 있는 것으로 알려져 있지만, 본 연구에서는 입력변수의 차이만으로 통계적으로 유의한 성능 변화를 이끌어낼 수 있음을 보이기 위하여 가중치 조정을 하지 않았다. 인공지능망 모형과 마찬가지로 모든 입력변수는 3개의 범주로 등분위 범주화(Equal frequency binning)하였다. 기존 연구에서 발췌한 39개의 입력변수에 대해 전처리 과정을 거치지 않고, 모든 변수를 입력변수로 사용하여 사례기반추론 모형을 생성하였다. 이들 39개의 변수는 기존 연구에서 기업의 신용등급을 결정하는데 유의한 것으로 나타난 것들이라 그런지, 본 실험에서도 비교적 그 적중률이 높게 나타났다.

판별분석의 stepwise 옵션을 이용하여 선정된 판별 모형의 독립변수들을 사례기반추론(CBR, Case-Based Reasoning)의 입력변수로 사용하였다. 이는 전통적인 기업신용평가 모형에서 판별 분석을 인공지능망 같은 분류 모형의 입력변수 선정을 위해 사용하는 것을 응용한 모형이라고 할 수 있다.

주성분-사례기반추론 모형은 러프집합 분석을 통해 생성된 코어와 코어에 근접하는 변수들을 사례기반추론의 입력변수로 사용한 것이다. 입력변수의 개수는 최초 1개에서 다음 순위의 코어를 하나씩 추가할 때마다 나타나는 적중률 분포의 추이를 보면서 적중률이 급격하게 감소하거나

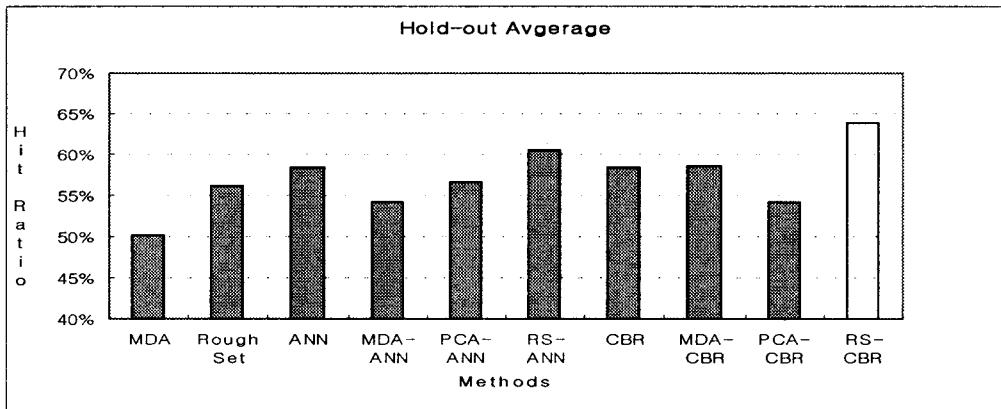


혹은 일정 수준 이상의 증감이 나타나지 않을 때로 결정하였다. 하지만 변수의 개수가 지나치게 많아지면 모형으로서의 의미가 없어지기 때문에 최대 20개를 넘지 않도록 하였다. 자료는 3등분 위로 범주화(Equal frequency binning) 하였고, 유전자 알고리즘(GA, Genetic Algorithm)을 이용하여 리덕트와 규칙을 생성하였다.

### 5.3 결과

#### 5.3.1 실험모형 결과 비교

<그림 3> 및 <표 5>는 본 연구에서 제시한 10개의 모형에 대한 모든 실험 별 적중률 분포를 나타낸 것이다. 전반적으로 판별 분석 또는 러프집합 분석만으로 구성된 단일 모형보다는 고유의 변수 선정 기법과 분류 모형(Classifier)을 통합한 결합 모형(Hybrid Model)에서 좀 더 우수한 성과를 보여주고 있다.



<그림 3> 10개 모형의 평균 적중률 분포

또한 결과에서 나타난 것처럼 변수 선정을 하지 않고 최초의 39개 변수를 모두 입력 변수로 사용한 인공신경망(ANN), 사례기반추론(CBR) 모형의 적중률도 상당히 높은 편이다. 이는 39개의 변수 모두가 기존의 연구에서 그 통계적 유의성이 검증된 것이므로 비교적 노이즈가 많이 포함되지 않아 그런 것으로 추정할 수 있다. 반면 판별 분석(MDA), 주성분 분석(PCA)를 이용한 변수 선정은 위 두 기법에 비해 비교적 적중률이 낮은 것으로 나타났다. 무엇보다 판별 분석에서 적중률이 낮은 것은 1차 전처리 과정에서 상관 분석 등을 통해 중요한 변수군이 제거된 것 등을 그 이유로 들 수 있다. 마지막으로 러프집합 분석을 통해 선정된 변수군은 인공신경망과 사례기반추론 어느 기반 하에서나 우수한 성과를 보이고 있다. 특히 사례기반추론 하에서 그 성능이 더욱 뛰어난 것으로 나타나는데, 이는 뒤에서 검증 결과를 제시하겠지만 이진 분류(Binary Classification)가 아닌 다분류(Multiple Classification) 문제에서 그리고 데이터에 오류가 많지 않

은 정확한 사례베이스를 구축하였을 경우에는 인공신경망보다 사례기반추론이 뛰어난 성과를 보일 수도 있음을 입증하는 것이다.

<표 5> 10개 모형의 실험 별 적중률

Experiment #	Data set	MDA	Rough Set	ANN	MDA-ANN	PCA-ANN	RS-ANN	CBR	MDA-CBR	PCA-CBR	RS-CBR
1	Training	59.10%	-	64.39%	58.32%	64.77%	61.03%	-	-	-	-
	Test		-	55.00%	53.08%	51.92%	59.23%	-	-	-	-
	Hold-out	53.96%	57.31%	60.33%	57.69%	56.92%	65.38%	62.31%	53.21%	57.89%	69.20%
2	Training	60.80%	-	66.50%	66.50%	61.94%	60.13%	-	-	-	-
	Test		-	63.46%	61.15%	53.08%	55.77%	-	-	-	-
	Hold-out	36.90%	56.54%	51.15%	43.46%	57.31%	59.46%	56.54%	57.31%	54.62%	56.90%
3	Training	57.30%	-	70.45%	62.06%	65.16%	67.48%	-	-	-	-
	Test		-	67.69%	61.15%	56.15%	67.69%	-	-	-	-
	Hold-out	52.70%	54.23%	63.09%	59.46%	57.31%	59.23%	59.08%	50.77%	51.92%	62.00%
4	Training	57.70%	-	62.97%	56.65%	63.87%	64.26%	-	-	-	-
	Test		-	59.23%	54.23%	50.77%	64.62%	-	-	-	-
	Hold-out	51.90%	56.92%	59.62%	55.38%	56.92%	59.09%	53.23%	59.08%	52.31%	64.90%
5	Training	59.90%	-	71.61%	63.35%	63.87%	63.48%	-	-	-	-
	Test		-	64.62%	59.62%	50.77%	59.23%	-	-	-	-
	Hold-out	55.80%	59.36%	57.31%	56.19%	54.62%	66.15%	55.77%	57.31%	54.23%	66.90%
Hold-out Avg.		50.16%	56.08%	58.31%	54.23%	56.62%	60.46%	58.38%	58.54%	54.15%	63.86%

### 5.3.2 유의성 검정

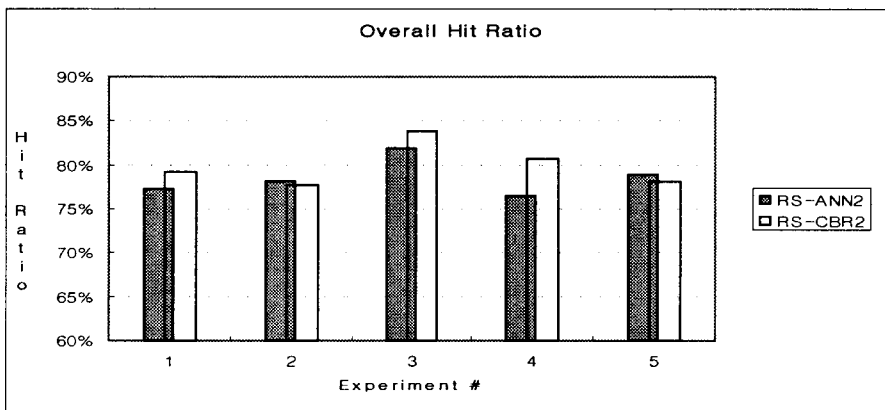
본 논문에서 제안한 기업신용평가 모형은 러프집합과 사례기반추론을 결합한 것으로, 실험 결과 적중률 분포에서 다른 통계적 기법 및 인공지능 기법에 비해 우수한 성과를 보이고 있다. 이러한 적중률의 차이가 과연 통계적으로도 의미가 있는 유의한 것인지를 검증하기 위하여 McNemar Test를 실시하였고 (<표6> 참조), 그 결과 러프집합-사례기반추론 (RS-CBR) 모형의 실험 결과는 다른 모형들과 통계적으로도 차이가 있는 의미 있는 것으로 결론지을 수 있었다.

<표 6> 모형 간 적중률 차이에 대한 McNemarTest

	Rough Set	ANN	MDA-ANN	PCA-ANN	RS-ANN	CBR	MDA-CBR	PCA-CBR	RS-CBR
MDA	6.5187**	25.4032***	6.4844**	11.1472***	37.6362***	22.2495***	23.1889***	4.8891**	58.0167***
Rough Set		58.6745***	28.8368***	40.7351***	78.9174***	49.8300***	54.5919***	22.7325***	99.2385***
ANN			10.2037***	0.8579	2.4795	0.0000	0.0084	5.5956**	12.3553***
MDA-ANN				1.8405	18.3381***	6.1872**	7.0678***	0.0000	34.8662***
PCA-ANN					4.4795**	0.8026	0.9948	1.8340	16.1362***
RS-ANN						1.4922	1.2075	12.666***	6.2465**
CBR							0.0024	4.9507**	10.4893***
MDA-CBR								5.6915**	9.3038***
PCA-CBR									90.8481***

주) \*\*\*-1% 수준에서 유의, \*\*-5% 수준에서 유의, \*-10% 수준에서 유의

일반적으로 분류 모형(Classifier)으로서의 사례기반추론은 변수간 중요도에 따른 가중치 설정을 하지 않는다면 인공신경망에 비해 그 성과가 높지 않은 것으로 알려져 있다. 하지만 본 논문에서 제안한 러프집합-사례기반추론(RS-CBR) 모형의 경우 그 적중률이 러프집합-인공신경망(RS-ANN) 모형보다 우월한 것으로 나타났다. 본 연구에서는 부도 예측이나 주가 방향성 예측과 같은 이진 분류(Binary Classification)가 아닌 다분류(Multiple Classification) 문제의 고유한 특성이 두 모형간 적중률에 유의한 차이를 일으켰을 것이라는 가설을 세워보았다. 이러한 가설을 검증하기 위해 본 모형 구축에 사용된 동일한 기업등급 데이터에 대해 상위 2개 등급(A1, A2)을 1개의 등급(0)으로, 그리고 하위 3개 등급(A3, B, C)을 나머지 1개의 등급(1)으로 이진 변환하여 그 성과를 모형 별로 비교하였다. <그림 4>는 모형 간 실험 별 적중률 분포를 나타내고 있다. 전반적으로 유사한 분포를 보이고 있고, McNemar Test에서도 통계적으로 유의한 차이가 나타나지 않았다. <표 7>에서 나타난 것과 같이 이진 분류한 데이터에 대해서는 두 모형 간 통계적으로 유의한 차이가 없음을 알 수 있다. 따라서 본 연구에서 제안한 러프집합-사례기반추론(RS-CBR) 모형은 이진 분류보다는 다분류 문제에서 그 성능이 더욱 뛰어난 것으로 판단할 수 있다.



<그림 4> 이진 분류 문제에 대한 모형 간 비교

<표 7> 이진 분류 문제에 대한 McNemar Test

	McNemar Value	p value
RS-ANN & RS-CBR	1.7410	0.1870

## VI. 결 론

본 논문의 목적은 기업신용평가라는 영역에서 사례기반추론의 적용 가능성과 사례기반추론을 위한 자료 및 변수 축약 기법으로서의 러프집합이론의 유용성을 알아보는 것이다. 이러한 연구 목적을 달성하기 위해 크게 세 가지의 연구 모형을 설계하였다. 첫째, 단일 모형으로서 판별 분석과 러프집합 분석을 이용한 모형이다. 이는 결합 모형(hybrid model)들과의 적중률 비교를 위한 기준을 제공하기 위해 설계되었다. 둘째, 인공신경망을 분류 모형(Classifier)으로 하는 세 가지 결합 모형이다. 이 세 가지는 각각 판별 분석, 주성분 분석, 러프집합 분석을 변수 전처리 과정으로 사용한 모형들이다. 마지막으로 위 세 가지 변수 전처리 기법과 분류 모형으로서의 사례기반추론을 결합한 모형이다.

실험 결과, 전반적으로 사례기반추론의 분류 능력이 인공신경망에 비해 우수함을 알 수 있었고, 러프집합 분석을 통해 선정된 입력변수들이 다른 통계 기법에서 선정된 변수들에 비해 입력 변수로서의 유용성이 뛰어난 것을 알 수 있었다.

본 연구를 통해 다음과 같은 의미를 찾을 수 있었다. 첫째, 기존 연구에서 분류 모형(Classifier)으로 널리 사용되는 인공신경망과 사례기반추론 두 개의 플랫폼 하에서 기업신용평가 모형을 구축한 결과, 사례기반추론이 인공신경망에 비해 우수한 성과를 나타내었다. 다만 주성분 분석을 변수 선정 기법으로 사용할 경우 인공신경망이 좀 더 우수한 성과를 보였지만 통계적으로 유의한 차이가 아닌 것으로 나타났다. 사례기반추론이 우수한 성과를 보이는 데는 여러 이유가 있을 수 있겠지만, 신용평가와 같은 다분류(Multiple Classification) 문제의 경우 인공신경망보다 뛰어난 성과를 보일 수 있음을 비교 실험을 통해 알 수 있었다. 즉, 동일한 기업자료에 대해 그 신용등급을 상(0), 하(1)로 이진 분류하여 실험한 결과 인공신경망과 사례기반추론 사이에 통계적으로 유의한 성과 차이가 나타나지 않았다. 둘째, 위 두 개의 분류 모형에 대해 다양한 변수 선정 기법을 적용한 결과 러프집합 분석을 이용할 때 다른 기법들에 비해 일관되게 뛰어난 성능을 보였다. 본 논문의 목적이 사례기반추론 하에서 러프집합이론의 자료 및 변수 축약 능력을 입증하려는 것임을 고려할 때 고무적인 결과라 할 수 있다.

본 연구에서 고려되지 못하였거나 실험 과정에서 문제점으로 첫째, 러프집합 분석과 비교하기 위해 사용된 변수 선정 기법들은 예외 없이 오픈-루프(open-loop) 방식을 따르는 방법들이었다. 물론 근본적으로 러프집합 분석이 오픈-루프 방식이므로 클로즈드-루프(closed-loop) 방식과 비교하는 것은 옳지 않다고 볼 수 있지만, 실무에서는 모형의 형식보다 예측률의 향상을 중요하게 생각하므로 유전자 알고리즘과 같은 클로즈드-루프 방식과의 성과 비교를 하지 못한 점은 분명 미흡하다고 할 수 있다. 둘째, 사례기반추론이 다분류(Multiple Classification) 문제에서 인공신경망에 비해 그 성과가 뛰어날 수 있음을 이진 분류화한 데이터에 대한 추가 실험을 통해서 간접적으로 증명하였지만, 이외에도 존재할 수 있는 여러 복합적인 요인에 대한 설명이 부족하였다. 물론 데이터 기반(data-driven) 방법론의 일종의 한계로 간주할 수도 있지만 모형의 총체적인 설득력을 높이기 위해서는 그 결과에 대한 충분한 설명도 반드시 뒤따라야 할 것이다. 따라서

사례기반추론과 러프집합 분석이라는 방법론에 대한 추가적인 실험 외에도 기업신용평가라는 분야 자체가 가지는 근본적 특성에 대한 이해가 요구된다.

본 논문에서 나타나는 한계점들을 고려하여, 다음의 과제를 향후의 연구로 계속하고자 한다.

첫째, 본 논문에서 제안한 방법론이 신용평가뿐 아니라 부도 예측이나 주가 방향성 예측과 같은 또 다른 경영 문제에 적용해도 일관되게 우수한 성능을 보여주는지 알아봐야 할 것이다. 이는 좀 더 일반화된 모형 개발의 시발점이 된다는 점에서 매우 중요한 과제이다.

둘째, 본 연구에서 보여주는 성과가 단지 오픈-루프(open-loop) 방식에 국한된 것인지 아니면 클로즈드-루프(closed-loop) 방식의 방법론들과 비교해도 우월한 성능을 보여주는 지 알아볼 필요가 있다. 이를 위해서는 우선 가장 대표적인 클로즈드-루프 방식의 변수 선정 기법인 유전자 알고리즘과 비교해봐야 할 것이다.

셋째, 러프집합 분석의 단점으로 모형에 따른 리덕트의 수가 지나치게 많이 생기는 점을 들 수 있는데, 이 때문에 모든 가용한 리덕트에 대해 그 적중률을 보는 것이 어려운 경우가 많다. 본 연구에서도 생성된 리덕트 각각을 입력변수군으로 하는 모형을 구축하고 실험하는 것이 쉽지 않아 그 차선책으로 코어(Core)를 선택하였다. 물론 코어 자체도 입력변수로서 매우 중요한 의미를 가지고 있지만 가능하다면 하나의 리덕트가 하나의 독립된 입력변수군을 형성하는 모형을 만들고 그 성과를 알아보는 것도 의미 있을 것이라 생각된다. 개별 리덕트를 입력변수군으로 사용하는 작업을 일일이 수작업에 의존하지 않고 인공지능망이나 사례기반추론 같은 분류 모형과 하나로 통합하여 자동화된 시스템을 구축하는 후속 작업이 필요하다.

마지막으로, 사례기반추론의 변수 별 가중치를 조정하여 더욱 성과가 뛰어난 모형을 개발하는 작업이 요구된다. 일반적으로 사례기반추론은 입력변수의 중요도에 따라 그 가중치를 차등 설정함으로써 모형의 적중률을 향상시킬 수 있는 것으로 알려져 있지만, 본 연구에서는 입력변수의 차이만으로 그러한 결과를 나타낼 수 있음을 보이기 위하여 가중치를 설정하지 않았다. 따라서 향후 사례기반추론의 분류 능력을 높이기 위해서 입력변수의 가중치를 합리적으로 조정하는 모형의 개발이 필요하다.

## 참 고 문 헌

- 김창연, 안병석, 조성식, 김성희, “도산 예측을 위한 러프집합이론과 인공지능망 통합방법론,” 경영정보학연구, Vol.9(4), 1999, pp.23-40.
- 박기남, 이훈영, 박상국, “러프집합을 이용한 통합형 채권등급 평가모형 구축에 관한 연구,” 한국경영과학회지, Vol.25(3), 2000, pp.125-135.
- 이전창, 한인구, 김명중, (1996). “통계적 모형과 인공지능 모형을 결합한 기업신용평가 모형에 관한 연구,” 한국경영과학회지, Vol.21(1), pp.81-100.
- Aamodt, A. and Plaza, E., “Case-based reasoning: Foundational issues, methodological

- variations, and system approaches,” *Artificial Intelligence Communications*, Vol.7(1), 1994, pp.39 - 59.
- Altman, E., “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *The Journal of Finance*, Vol.23(4), 1968. pp.589-609.
- Altman, E., *A Complete Guide to Predicting, Avoiding and Dealing with Bankruptcy*, John Wiley & Sons, New York, 1983.
- Barbro, B., Tejja, L. and Kaisa, S., “Neural networks and genetic algorithms for bankruptcy predictions,” *Proceedings of the Third World Congress on Expert Systems*, 1996, pp.123-130.
- Barletta, R., “An introduction to case-based reasoning,” *AI Expert*, Vol.6 (8), 1991, pp.42-49.
- Beaver, W., “Financial ratios as predictors of failure,” *Empirical Research in Accounting: Selected Studies, Supplement to Journal of Accounting Research*, Vol.4, 1996, pp.71-111.
- Brown, C.E., and Gupta, U.G., “Applying case-based reasoning to the accounting domain.” *Intelligent Systems in Accounting, Finance and Management*, Vol.3, 1994, pp.205-221.
- Cadden, D.T., “Neural network and mathematics of chaos - An investigation of these methodologies as accurate predictors of corporate bankruptcy,” *Proceedings of the First International Conference on Artificial Intelligence Application on Wall Street*, 1991, pp.52-57.
- Chung, H. and Tam, K., “A comparative analysis of inductive learning algorithm,” *Intelligent Systems in Accounting, Finance and Management*, Vol.2, 1992, pp.3-18.
- Dimitras, A.I., Slowinski, R., Susmaga, R. and Zopounidis, C., “Business failure prediction using rough sets,” *European journal of operational research*, Vol.114(2), 1999, pp.263-280.
- Elmer, P.J. and Borowski, D.M., “An expert system approach to financial analysis: The case of S&L bankruptcy,” *Financial Management, Autumn*, 1988, pp.67-76.
- Gentry, J.A. Newbold, P., and Whitford, D.T., “Classifying bankrupt firms with funds flow components,” *Journal of Accounting Research, Spring*, 1985, pp.146-160.
- Hammer, M., “Failure prediction: Sensitivity of classification accuracy to alternative statistical method and variable sets,” *Journal of Accounting and Public Policy*, 1983, pp.289-307.
- Hashemi, R.R., Le Blanc, L.A., Rucks, C.T. and Rajaratnam, A., “A hybrid intelligent system for predicting bank holding structures,” *European Journal of Operational Research*, Vol.109, 1998, pp.390-402.
- Huang, C.C. and Tseng, T.L., “Rough set approach to case-based reasoning application,” *Expert Systems with Applications*, Vol.26(3), 2004, pp.369-385.
- Jo, H., Han, I. and Lee, H., “Bankruptcy prediction using case base reasoning, neural

- network, and discriminant analysis," *Expert Systems with Applications*, Vol.13(2), 1997, pp.97-108.
- Johnson, W.B., "The cross-sectional stability of financial ratio patterns," *Journal of Financial and Quantitative Analysis*, Vol.14, 1979, pp.1035-1048.
- Kim, K. and Han, I., "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," *Expert Systems with Applications*, Vol.19(2), 2000, pp.125-132.
- Kingdom, J. and Feldman, K., "Genetic algorithms for bankruptcy prediction," *Search Space Research Report*, No.1. 1995.
- Kolodner, J.L., *Case-based reasoning*, Morgan Kaufmann: Los Altos, CA, 1993.
- Lee, K.C. and Kim, J., "Hybrid neural network-driven reasoning approach to bankruptcy prediction: Comparison with MDA, ACLS, and neural network," *Proceedings of the 1994 IEEE International Conference on Neural Networks*, 1994, pp.1787-1792.
- Lee, S.B. and Oh, S.H., "A comparative study of recursive partitioning algorithm and analogue concept learning system," *Expert Systems with Applications*, Vol.1, 1990, pp.403-416.
- Martin, D., "Early warning of bank failure: A logit regression approach," *Journal of Banking and Finance*, Vol.1, 1977, pp.249-276.
- McKee, T.E. and Lensberg, T., "Genetic programming and rough sets: A hybrid approach to bankruptcy classification," *European Journal of Operational Research*, Vol.138, 2002, pp.436-451.
- Miller, W., Cadden, D.T. and Driscoll, V., "Bank failure and categorization - A neural network approach," *Proceedings of the Third International Conference on Artificial Intelligence Applications on Wall Street*, 1995, pp.232-235.
- Odom, M. and Sharda, R., "A neural networks model for bankruptcy prediction," *Proceedings of the IEEE International Conference on Neural Network*, Vol.3, 1990, pp.163-168.
- Ohlson, J., "Financial ratios and probabilistic prediction of bankruptcy," *Journal of Accounting Research*, Vol.18(1), 1980, pp.109-131.
- Pawlak, Z., "Rough sets," *International Journal of Information and Computer Science*, Vol.11, 1982, pp.314-356.
- Pawlak, Z. and Sowinski, R., "Rough set approach to multi attribute decision analysis," *European Journal of Operational Research*, Vol.72(3), 1994, pp.443-459.
- Pawlak, Z., "Rough set approach to knowledge-based decision support," *European Journal of Operational Research*, Vol.99(1), 1997, pp.48-57.
- Riesbeck, C.K. and Schank, R.C., *Inside case-based reasoning*, Lawrence Erlbaum Associates, NJ:Hillsdale, 1989.

- Shin, K.S. and Han, I., "Case-based reasoning supported by genetic algorithms for corporate bond rating," *Expert Systems with Applications*, Vol.16, 1999, pp.85-95.
- Shin, K.S., Hong, S.H., and Lee, S.E., "A corporate failure prediction model using temporal pattern recognition," *Institute for Operation Research and the Management Sciences*, Seoul, Korea, 2000.
- Slade, S., "Case-based reasoning: A research paradigm," *AI Magazine*, Vol.12(1), 1991, pp.42-55.
- Slowinski, R. and Zopounidis, C., "Application of the rough set approach to evaluation of bankruptcy risk," *Intelligent Systems in Accounting, Finance and Management*, Vol.4, 1995, pp.27-41.
- Slowinski, R., Zopounidis, C., and Dimitras, A.I., "Prediction of company acquisition in Greece by means of the rough set approach," *European Journal of Operational Research*, Vol.100(1), 1997, pp.1-15.
- Swiniarski, R.W. and Hargis, L., "Rough sets as a front end of neural-networks texture classifiers," *Neurocomputing*, Vol.36, 2001, pp.85-102.
- Takahashi, K., Kurokawa, Y., and Watase, K., "Corporate bankruptcy prediction in Japan," *Journal of Banking and Finance*, 1984, pp.229-247.
- Tam, K. and Kinag, M., "Managerial applications of neural networks: the case of bank failure predictions," *Management Science*, Vol.38(7), 1992, pp.926-947.
- Yang, A.R., Platt, M.B., and Platt, H.D., "Probabilistic neural networks in bankruptcy prediction," *Journal of Banking and Finance*, Spring, 1999, pp.19-45.
- Zmijewski, M.E., "Methodological issues related to the estimation of financial distress prediction models," *Journal of Accounting Research*, Vol.22, 1984, pp.59-82.



<Abstract>

## Integrating rough set theory and case-based reasoning for the corporate credit evaluation

Tae-Hyup Roh, Myung-Hwan Yoo, In-Goo Han

The credit rating is a significant area of financial management which is of major interest to practitioners, financial and credit analysts. The components of credit rating are identified and decision models are developed to assess credit rating and the corresponding creditworthiness of firms as accurately as possible. Although many early studies demonstrate that one technique outperforms the others for a given data set, there is often no way to tell a priori which of these techniques will be most effective to solve a specific classification problem. Recently, a number of studies have demonstrated that a hybrid model integrating artificial intelligence approaches with other feature selection algorithms can be alternative methodologies for business classification problems.

In this article, we propose a hybrid approach using rough set theory as an alternative methodology to select appropriate attributes for case-based reasoning. This model uses rough set theory to extract knowledge that can guide effective retrievals of useful cases. Our specific interest lies in the stable combining of both rough set theory and case-based reasoning in the problem of corporate credit rating. In addition, we summarize backgrounds of applying integrated model in the field of corporate credit rating with a brief description of various credit rating methodologies.

**Keywords:** Corporate Credit Evaluation, Rough Set Theory, Cased-Based Reasoning

\* 이 논문은 2005년 3월 11일 접수하여 1차 수정을 거쳐 2005년 4월 18일 게재 확정되었습니다.