지능정보연구 제16권 제3호 2010년 9월(pp.77~97)

러프집합분석을 이용한 매매시점 결정

허진영
㈜네오아이즈
jinnyong.huh@neoiz.co.kr

김경재
동국대학교_서울 경영정보학과
kjkim@dongguk.edu

한인구
한국과학기술원 경영대학
ighan@business.kaist.ac.kr

매매시점결정은 금융시장에서 초과수익을 얻기 위해 사용되는 투자전략이다. 일반적으로, 매매시점 결정은 거래를 통한 초과수익을 얻기 위해 언제 매매할 것인지를 결정하는 것을 의미한다. 몇몇 연구자들은 러프집합분석이 매매시점결정에 적합한 도구라고 주장하였는데, 그 이유는 이 분석방법이 통제함수를 이용하여 시장의 패턴이 불확실할 때에는 거래를 위한 신호를 생성하지 않는다는 점 때문이었다. 러프집합은 분석을 위해 범주형 데이터만을 이용하므로, 분석에 사용되는 데이터는 연속형의 수치값을 이산화하여야 한다. 이산화란 연속형 수치값의 범주화 구간을 결정하기 위한 적절한 "경계값"을 찾는 것이다. 각각의 구간 내에서의 모든 값은 같은 값으로 변환된다. 일반적으로, 러프집합분석에서의 데이터 이산화 방법은 등분위 이산화, 전문가 지식에 의한 이산화, 최소 엔트로피 기준 이산화, Naïve and Boolean reasoning 이산화 등의 네 가지로 구분된다. 등분위 이산화는 구간의 수를 고정하고 각 변수의 히스토그램을 확인한 후, 각각의 구간에 같은 숫자의 표본이 배정되도록 경계값을 결정한다. 전문가 지식에 의한 이산화는 전문가와의 인터뷰 또는 선행연구 조사를 통해 얻어진 해당 분야 전문가의 지식에 따라 경계값을 정한다. 최소 엔트로피 기준 이산화는 각 범주의 엔트로피 측정값이 최적화 되도록 각 변수의 값을 재귀분할 하는 방식으로 알고리즘을 진행한다. Naïve and Boolean reasoning 이산화는 Naïve scaling 후에 그로 인해 분할된 범주값을 Boolean reasoning 방법으로 종속변수 값에 대해 최적화된 이산화 경계값을 구하는 방법이다. 비록 러프집합분석이 매매시점결정에 유망할 것으로 판단되지만, 러프집합분석을 이용한 거래를 통한 성과에 미치는 여러 이산화 방법의 효과에 대한 연구는 거의 이루어지지 않았다. 본 연구에서는 러프집합분석을 이용한 주식시장 매매시점결정 모형을 구성함에 있어서 다양한 이산화 방법론을 비교할 것이다. 연구에 사용된 데이터는 1996년 5월부터 1998년 10월까지의 KOSPI 200데이터이다. KOSPI 200은 한국 주식시장에서 최초의 파생상품인 KOSPI 200 선물의 기저 지수이다. KOSPI 200은 제조업, 건설업, 통신업, 전기와 가스업, 유통과 서비스업, 금융업 등에서 유동성과 해당 산업 내의 위상 등을 기준으로 선택된 200개 주식으로 구성된 시장가치 가중지수이다. 표본의 총 개수는 660거래일이다. 또한, 본 연구에서는 유명한 기술적 지표를 독립변수로 사용한다. 실험 결과, 학습용 표본에서는 Naïve and Boolean reasoning 이산화 방법이 가장 수익성이 높았으나, 검증용 표본에서는 전문가 지식에 의한 이산화가 가장 수익성이 높은 방법이었다. 또한, 전문가 지식에 의한 이산화가 학습용과 검증용 데이터 모두에서 안정적인 성과를 나타내었다. 본 연구에서는 러프집합분석과 의사결정나무분석의 비교도 수행하였으며, 의사결정나무분석은 C4.5를 이용하였다. 실험결과, 전문가 지식에 의한 이산화를 이용한 러프집합분석이 C4.5보다 수익성이 높은 매매규칙을 생성하는 것으로 나타났다.

논문접수일 : 2010년 06월 15일 논문수정일 : 2010년 07월 10일 게재확정일 : 2010년 7월 21일 주저자 : 김경재

1. 서론

데이터마이닝 기술의 한 부류인 러프집합은 Pawlak (1982)에 의해 제안된 이래 많은 연구자들에 의해 이론적 발전을 이루어 왔다. 러프집합은 데이터로부터 일정한 패턴을 보이는 규칙을 비교적 용이하게 추출할 수 있다는 것이 가장 큰 장점이라고 할 수 있다. 이러한 장점을 이용하여 경영학 분야에서

도 많은 선행연구들이 러프집합을 이용하였다. 러프집합을 응용한 대표적인 경영학 분야로는 기업부실예측과 주가지수 예측 등이 있다(Tay and Shen, 2002).

러프집합은 많은 장점을 지닌 데이터마이닝 방법의 하나이지만 실제 러프집합을 이용하여 응용연구를 수행하기 위해서는 여러 가지 어려움을 겪게 된다. 대표적인 어려움으로는 러프집합에서 생성된 정보 테이블(information table)로부터 의사결정규칙을 추출할 때, 의사결정에 유의한 의사결정규칙만을 어떻게 추출할 것인가의 문제이다. 하나의 정보테이블로부터 얻을 수 있는 의사결정규칙은 매우 많으나 일반화된 의사결정규칙만을 도출하기 위해서는 얻어진 의사결정 규칙 중에서 유의한 것만을 추출해 내야 한다는 것으로, 이 문제에 대한 해결방법은 Grzymala-Busse(1992), Skowron(1993), Slowinski and Stefanowski(1994) 등에 의해 여러 가지 대안이 제시된 바 있다.

한편, 주식시장 자료와 같은 재무자료에 대해 러프집합분석을 응용할 때에는 또 다른 어려움이 있는데, 이는 연속형 자료의 이산화(discretization)에 관련된 문제이다. 러프집합분석은 일반적으로 일정한 구간으로 나누어진 속성값을 갖는 이산화된 자료를 대상으로 분석을 할 수 있는데, 주식시장 자료와 같은 재무자료는 일반적으로 자료형태가 연속형으로 구성되는 경우가 많으며, 이러한 형태로는 러프집합이 의사결정규칙을 생성할 수 없으므로 연속형 자료를 이산형의 자료로 변환시켜주는 과정이 필요하다. 그러나, 이산화 과정을 수행하는 방법에 대해서는 일반적으로 인정된 원칙이나 방법이 없다. 또한 이에 대한 선행연구도 거의 없다. 그러나 어떠한 이산화 방법을 사용하느냐에 따라 분석의 결과는 매우 큰 차이가 발생할 수 있으므로 적절한 이산화 방법의 개발이 필요하다.

본 연구에서는 주식시장 자료와 같이 연속형 자료를 다수 포함한 자료를 러프집합을 이용하여 분석할 때 사용할 수 있는 여러 가지 이산화 방법들을 제시하고 각 방법의 성과를 비교하여 러프집합을 이용한 재무자료 분석에 적합한 이산화 방법론을 제시하고자 한다.

본 연구는 다음과 같이 구성되어 있다. 다음 장에서는 선행연구들을 정리하고 러프집합이론의 기본 개념과 이를 경영문제에 응용한 연구들을 살펴본다. 제 3장에서는 러프집합분석을 실행할 수 있는 실험설계에 대해 설명하고, 제 4장에서는 주식시장에서의 매매시점 결정에 사용될 규칙을 추출하기 위한 방법을 제시하고 실험에 사용될 자료와 실험방법을 기술한다. 제 5장에서는 주식시장에서의 거래 모의실험 결과를 제시하고 제 6장에서는 연구의 결론과 향후 연구과제를 제안한다.

2. 연구배경

2.1 러프집합

러프집합은 Pawlak(1982)에 의해 처음 제시되었다. 이 이론은 불확실한 자료를 용이하게 처리할 수 있다는 점에서 확률이론, 증거이론, 그리고 퍼지이론과 유사성을 가진다. 기본적으로 러프집합은 세상의 모든 개체들이 그들이 가진 특정한 정보로써 집합을 만들 수 있다는 가정 하에서 시작된다. 동일한 정보의 외연을 가진 개체들은 동일한 것으로 취급되며, 이러한 동질성 관계(indiscernibility relationship)가 러프집합이론의 기초가 된다. 동질성을 가진 집합을 러프집합이론에서는 기본집합(elementary set)이라 하고 이들이 하나의 전체집합에 대한 지식의 기본 단위를 이룬다. 또한, 특정 기본집합의 합집합이 되는 집합을 일반집합(crisp set)이라 하고 그렇지

않은 경우를 러프집합(rough set)이라 한다.

일반적으로 러프집합은 하위근사(lower approximation)와 상위근사(upper approximation)로 불리는 일반 집합들로 표현될 수 있다. 전자는 어떤 개체가 소속하고자 하는 집합에 확실성을 가지고 소속되는 경우이고, 후자는 소속하고자 하는 집합에 속할 수도 있고 속하지 않을 수도 있는 경우라고 할 수 있다.

러프집합에서 개체들의 정보는 정보테이블의 형태로 표현된다. 정보테이블은 기본적으로 행과 열로 구성되는데, 각 행은 하나의 개체를 표현하며, 열은 그 개체들에 대한 속성들로 이루어진다. 따라서 그들이 교차하는 셀 부분이 개체들의 속성값으로 채워져 있는 테이블 형태로 나타낼 수 있다. 정보테이블 S를 수식의 형태로 표현하면 식 (1)과 같다(Pawlak, 1997).

$$S = < U, Q, V, f > \quad (1)$$

S : 정보테이블

U : 개체집합(a finite sets of universe)

Q : 속성집합(a finite sets of attributes)

$V = \bigcup_{q \in Q} V_q$

$V_q = q$속성의 값의 집합

$f : U \times Q \rightarrow V$

정보테이블 S에서 전체 속성집합 Q의 부분집합 $P(\subseteq Q)$를 정의하고, 임의의 개체 $x, y(\in U)$가 존재할 때, $\forall q \in Q, f(x,q) = f(y,q)$의 조건을 만족하면 x, y는 어떤 속성집합 P에 대하여 동질성 관계를 가진다고 말한다. 집합 $P(\subseteq Q)$가 전체 개체집합에 대해서 쌍방의 동질성 관계를 형성시킬 때 이를 'P에 대한 동질성 관계(P-Indiscernibility relationship)'라 하고 I_P로 표현한다. 또한 전체 속성집합 Q 중에서 부분집합 $P(\subseteq Q)$를 뽑고, 전체 개체집합 U 중에서 부분집합 $Y(\subseteq Q)$를 선택하고, 집합 P에 대한 하위 근사와 상위근사를 각각 $\underline{P}Y$와 $\overline{P}Y$로 표현하면 이들을 다음 식 (2)와 같이 표현할 수 있다(Slowinski and Zopounidis, 1995).

$$\underline{P}Y = \{x \in Y, I_P(x) \subseteq Y\}, \ \overline{P}Y = \bigcup_{x \in Y} I_P(x) \quad (2)$$

또 다른 개념으로서 집합 Y의 집합 P에 대한 경계부분(P-Boundary of set Y)이라는 개념이 있는데, 이는 상위 근사집합에서 하위 근사집합에 속하지 못하는, 즉 확실성이 없는 부분을 의미하며, 이를 $BN_p(Y)$로 나타내고 그 의미는 아래 식 (3)과 같이 나타낼 수 있다.

$$BN_p(Y) = \overline{P}Y - \underline{P}Y \quad (3)$$

즉, 하위근사집합 $\underline{P}Y$는 P라는 속성집합에 의해 확실성 있게 Y집합에 속할 수 있는 전체 개체집합 U의 모든 원소들의 집합이라 할 수 있고, 상위근사집합 $\overline{P}Y$는 P라는 속성집합에 의해 Y집합에 속할 가능성이 있는 집합이라 할 수 있다. 따라서 상위 근사집합 $\overline{P}Y$와 하위 근사집합 $\underline{P}Y$의 차이인 $BN_p(Y)$는 확실성을 가지고 Y집합에 속할 수 없는 집합, 즉 Y집합에 속할 가능성이 있는 집합이라고 할 수 있다.

한편, 전체 개체집합의 부분집합이 되는 $Y(\subseteq U)$에서 어떤 속성집합 P를 이용하여 '정확도(accuracy of approximation)'라는 것을 정의할 수 있고, 이는 식 (4)에 의해 계산할 수 있다.

$$a_p(Y) = \frac{Cardinality(\underline{P}Y)}{Cardinality(\overline{P}Y)} \quad (4)$$

($Cardinality(x)$: x에 해당하는 개체의 개수)

즉, 정확도는 상위근사집합에 속하는 개체의 개수에 대한 하위근사집합에 속하는 개체의 개수의 비율이라고 할 수 있다. 이는 보다 쉽게 표현하면, 집합 Y에 확실하게 소속될 수 있는 개체의 개수와 집합 Y에 속할 가능성이 있는 개체 개수의 비율이라고 할 수 있다.

또한, 특정 정보테이블 S에서 전체 속성집합 Q의 어떤 부분집합 $P(\subseteq Q)$와 전체집합을 분할한 집합 $\mathscr{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ 를 가정한다. 이를 통해 속성집합 P에 의해 정의되는 집합 $\mathscr{Y}$에 대한 '근사정도(quality of approximation)'라는 개념을 식(5)와 같이 정의할 수 있다.

$$\gamma_P(\mathscr{Y}) = \frac{\sum_{i=1}^{n} Cardinality(\underline{P}Y_i)}{Cardinality(U)} \quad (5)$$

이는 정보테이블 내에서 속성집합 P가 의사결정 속성 클래스를 얼마나 잘 구분하는가를 나타내는 지표이다. 이 지표를 기준으로 유의한 속성군을 선택할 수 있다. 줄어든 속성군 $P(\subseteq Q)$가 전체속성군 Q가 가질 수 있는 분류와 같은 정도의 근사정도를 제공할 수 있다면 효율적인 분석을 위해 속성군의 수를 줄여 갈 수 있다. 여기서 같은 정도의 근사정도를 보장하면서, 가장 적은 속성의 수를 가진 속성군 $R(\subseteq P \subseteq Q)$을 '속성군 P의 $\mathscr{Y}$ - Reduct', 또는 'reduct'라 하고, 이를 '$RED_{\mathscr{Y}}(P)$'로 표현한다. 일반적으로 하나의 정보테이블은 하나 이상의 reduct를 가진다. 또한, 하나의 정보테이블에 여러 개의 reduct가 존재할 때, 모든 reduct의 교집합이 되는 속성들의 집합을 '속성군 P의 $\mathscr{Y}$ - Core', 또는 'core'라 하고, 이를 '$CORE_{\mathscr{Y}}(P)$'라고 표현한다.

이를 보다 쉽게 설명하면, 우리가 가진 분석자료에 여러 개의 속성이 있고, 이 중에서 선택된 소수의 속성만으로도 전체 속성을 사용하여 분석한 정도의 분석결과를 얻을 수 있다면 분석의 효율성을 위해 선택된 소수의 속성만을 이용하여 분석하는 것을 의미하며, 이 때 선택된 소수의 속성군을 reduct라 표현하고, 이런 reduct가 여러 개 있을 때, 각 reduct에 속한 속성들 중 공통적인 속성들을 core라 하는 것이다.

이상의 절차에 따라 러프집합분석을 적용하면 정보테이블에서 의사결정규칙을 생성시킬 수 있다. 의사결정규칙을 생성해내는 알고리즘은 다음과 같이 정리할 수 있다(Grzymala-Busse, 1992; Skowron, 1993).

(a) 정보 테이블의 모든 개체를 포괄할 수 있는 가장 적은 수의 규칙을 생성
(b) 정보 테이블을 위한 모든 가능한 규칙을 생성
(c) 정보 테이블의 모든 개체를 모두 다 포괄하지 않고 다수의 개체만을 포괄하지만 지지도가 높은 규칙을 생성

상기의 방법들에 따라 여러 개의 의사결정규칙이 생성될 수 있다. 그런데 의사결정규칙을 무한히 허용하게 되면 전체개체 수만큼의 의사결정규칙이 생성될 수도 있는데 이런 경우에는 향후에 일반화된 의사결정을 할 수 없게 된다. 따라서 생성된 의사결정규칙들 중에서 최종적으로 사용할 소수의 의사결정규칙을 선택하여야 하는데, 이 선택의 기준으로는 지지도(support 또는 strength)와 판별력(level of discrimination)을 사용한다(Dimitras et al., 1999). 지지도는 전체 개체 중에서 특정한 규칙을 지원하는 개체의 수를 의미하고, 판별력은 특정한 규칙에 의해 분류된 개체에 대한 클래스 분포를 나타내는 개념이다. 즉 지지도는 특정 의사결

정규칙이 적용될 수 있는 개체의 수를 의미하고, 판별력은 특정 의사결정규칙이 적용된 사례에서의 분류정확도를 의미하는 것이다.

일반적으로 의사결정규칙은 다수개가 생성되며 생성된 의사결정규칙을 새로운 개체에 적용할 때에는 특정 개체와 특정 규칙 간에 다음과 같은 4가지 상황이 발생할 수 있다(Slowinski and Stefanowski, 1994).

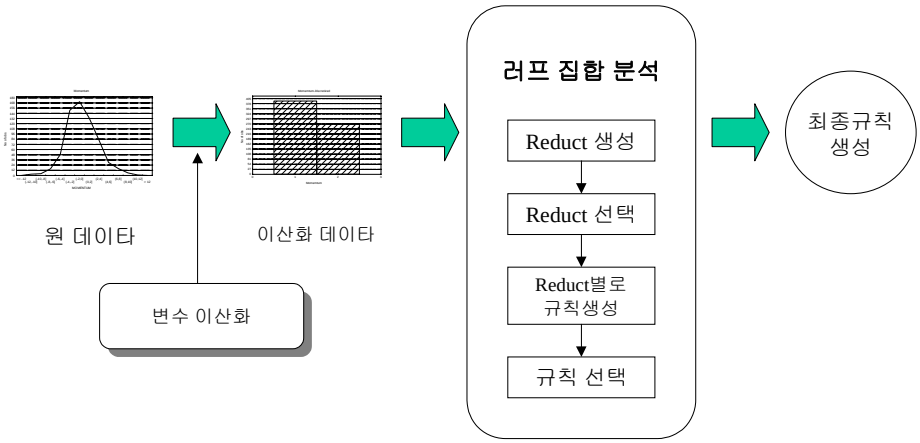
(a) 새로운 개체가 하나의 규칙에 정확하게 일치하는 경우
(b) 새로운 개체에 하나 이상의 규칙을 적용할 수 있는데, 그 규칙들의 결과 값이 모두 동일한 경우
(c) 새로운 개체에 하나 이상의 규칙을 적용할 수 있는데, 그 규칙들의 결과 값이 상이한 경우
(d) 새로운 개체에 적용할 수 있는 규칙이 존재하지 않는 경우

위에서 (a), (b)의 경우에는 하나 또는 여러 개의 의사결정규칙의 결과 값이 동일하므로 규칙을 적용하여 의사결정을 하는데 어려움이 없고, 따라서 해당 규칙을 적용하고 규칙의 결과 값을 산출하면 된다. 그러나 (c)의 경우에는 적용 가능한 각 규칙의 결과 값이 상이하므로 적용 가능한 규칙 중에서 개체에 적용할 규칙을 선택해야 하는 문제가 있다. 이러한 경우에 대해서 Dimitras et al.(1999)은 각 규칙의 지지도를 이용하여 적용우선순위를 정하는 방법을 사용하였다. (d)의 경우에는 특정 개체와 가장 유사한 형태의 규칙을 적용하거나, 또는 특정 개체에 대해서 규칙을 적용하지 않는 방법을 사용할 수 있다. 즉, 특정 사례와 유사한 다른 사례에 적용되는 규칙을 준용하여 이용하거나 아예 의사결정을 유보할 수 있다.

(d)의 경우에 있어서, 분석에 이용되는 모든 개체에 대해서 의사결정을 하여야 하는 분야의 문제, 예를 들면 기업신용평가와 같이 모든 기업에 대해 신용평가 의사결정을 하여야 하는 경우에 대해서 새로운 개체에 적합한 규칙이 존재하지 않는다면 새로운 개체와 가장 유사한 형태의 규칙을 적용하는 전자의 방법을 적용하여야 한다. 그러나, 모든 개체에 대한 의사결정을 필요로 하지 않는 분야의 문제, 예를 들면 주식시장분석에서의 매매시점 결정과 같이 모든 개체(거래일)에 대해 의사결정을 하지 않아도 무방한 경우에는 적용할 규칙이 적당하지 않은 (d)의 경우에는 의사결정을 유보하는 방법도 의미를 가질 수 있다. 주식거래는 주식을 계속 보유하거나 새로운 주식에 투자하지 않는 등의 방법으로 이용함으로써 특정 개체, 즉 특정 거래일에 대한 투자의사결정을 유보하는 방법을 이용할 수 있기 때문이다.

특히 러프집합분석은 다음과 같은 이유로 주가지수 선물시장에서 매매시점 포착을 위한 거래규칙을 찾아내는데 유용한 방법이다. 첫째, 러프집합분석은 의사결정규칙을 생성하는 과정에서 reduct와 core 등의 개념을 이용함으로써 분석자료 내에 포함되어 있는 잡음의 통제가 용이하고 예측과정에 불필요한 요소들을 제거하여 매매시점 포착을 위한 수익성 있는 거래 규칙을 발생시켜 줄 수 있다(Ruggiero, 1997). 둘째, 러프집합분석은 지지도와 판별력의 개념을 이용하여 예측이 불확실한 시장의 패턴에 대해서는 매매신호를 발생시키지 않도록 통제할 수 있는 장점이 있다. 셋째, 러프집합분석은 자료의 분포에 대한 특별한 가정을 요구하지 않기에 활용이 용이하다.

그런데 서론에서 언급한 바와 같이 연속형 자료에 러프집합을 적용하는 경우에는 연속형 자료를 규칙이 생성될 수 있는 이산형 자료로 변환시켜



<그림 1> 러프집합분석 적용 과정

주는 과정이 필요하다. 그러나 일반적인 러프집합 분석에서는 알고리듬 자체에 이산화를 수행하는 방법이 포함되어 있지 않을 뿐만 아니라 기존 연구에서도 이에 관한 일반화된 방법이 제시되지 못하였다. 선행연구에 따르면, 이산화 방법이 분류성과에 크게 영향을 미칠 가능성이 있기에 재무자료와 같이 연속형 자료가 대부분인 연구분야에서는 적절한 이산화 방법의 개발이 대단히 중요한 문제이다. 따라서 러프집합분석에서 일반적으로 활용될 수 있는 이산화 방법에 대한 검토가 반드시 필요하다고 할 수 있다.

이상의 절차에 따른 러프집합분석 적용의 일반적인 절차는 <그림 1>과 같다.

본 연구에서는 <그림 1>에서 제시된 일반적인 절차에 따라 주식시장에서의 매매시점 결정을 위한 의사결정규칙 생성에 러프집합분석을 적용하며, 여러 가지 이산화 방법의 효과를 검토할 것이다.

2.2 러프집합분석의 경영학 응용연구

비교적 최근에 경영학 분야에 적용된 러프집합 응용연구는 다음과 같다.

Slowinski and Zopounidis(1995)는 기업신용평가에 러프집합분석을 이용하였고, Slowinski et al.(1997)은 기업인수 가능성 평가에 러프집합분석을 응용하고 이의 성과를 다변량 판별분석과 비교하였다. 그리고 Dimitras et al.(1999)은 기업부실예측에 러프집합분석을 이용하고 그 결과를 로짓모형과 비교하였다. 이외에 다른 인공지능기법과 통합적으로 모형을 구축하고자 하는 노력 또한 있었는데 Hashemi et al.(1998)은 은행의 지배구조 예측문제에서 인공신경망의 전처리 과정으로써 러프집합분석을 이용하였다. 이 연구에서는 '2차원 감축(2-Dimensional reduction)'을 통해 최적 샘플과 속성으로 정보 테이블을 구성한 후 인공신경망을 이용하여 실험하였다. 김창연 등(1999)과 Ahn et al.(2000)은 Hashemi et al.(1998)과 유사한 방법론을 기업부실예측에 적용하였으며, 박기남 등(2000)은 채권등급평가 문제에 있어서 판별분석과 인공신경망의 결과값을 결합하는 방법으로 러프집합을 이용하였다. Kim and Han(2001)은 러프집합분석을 이용한 주식시장에서의 거래규칙 생성방법을 제안하였으며, 정석훈, 서용무(2008)는 러프집

합분석을 이용하여 신용카드 연체자를 분류하는 모형을 제안하였다. 이 연구에서는 러프집합분석의 성능이 입렵 변수의 수와 변수 범주화 방법에 매우 민감하게 반응하는 것으로 나타났다. Yeh et al.(2010)은 기업의 부실을 예측하기 위한 도구로써 SVM을 사용하였는데, 이의 입력변수 선정과정에 러프집합분석을 활용하여 정보손실 없이 입력변수를 최적화할 수 있었다고 보고하였다.

러프집합을 경영학 문제에 적용한 기존 연구는 전술한 바와 같이 주로 신용평가 등 기업평가와 관련된 분야에 집중되어 있다. 그러나, 기업평가 분야는 모든 새로운 개체(개별 기업)에 대하여 신용평가결정을 내려줄 필요가 있는데 이 경우에 러프집합분석을 적용하면 분석 후 산출되는 의사결정규칙이 과다하여 의사결정의 결과가 일반화되기 어려운 단점이 있다. 따라서 러프집합의 경우에는 모든 개체에 대해 의사결정을 할 필요가 없는 분야에서 더 유용하게 사용될 수 있다. 이는 러프집합에서 발생되는 의사결정규칙 중에서 지지도와 판별력 등의 기준을 이용하여 분류확실성이 높은 규칙만을 선택하여 사용할 수 있기 때문이다.

한편, 박기남 등(2000)의 연구에서는 판별분석과 인공신경망의 결과값, 즉 이산화된 채권등급 값을 이용하였으므로 러프집합을 실행하는데 있어서 별도의 이산화과정이 필요 없었다. 그러나 다른 연구들의 경우에는 대부분이 연속형으로 구성되어 있는 재무자료를 이용하여 분석하였으므로 이산화 과정이 반드시 필요하였다. 선행 연구들은 대부분 재무분야의 전문가의 의견을 반영하여 이산화의 기준을 정하고 이를 사용하여 연속형 자료를 이산화하여 사용하였다. 전문가의 의견을 이용하는 방법은 전문가의 지식을 분석과정에 반영할 수 있다는 장점이 있으나, 기준설정과정이 주관적이기 때문에 어떤 전문가가 참여하느냐에 따라 결과가 달라질 수 있다는 단점이 있다. 또한, 최근의 연구에서는 전문가의 지식 외에 다양한 방법을 활용한 이산화 방법이 소개되고 있다. 따라서 주관적인 방법 이외에 객관적으로 이산화 할 수 있는 방법을 비교 분석하여 재무분야에 러프집합을 적용할 때 적합한 이산화 방법을 제시하는 연구가 필요하다.

2.3 변수 이산화방법론

변수 이산화 방법은 연속형 값을 갖는 자료를 이산형 값을 갖는 자료로 변환해 주는 방법을 의미하며 변수 변환의 방법으로 많이 쓰이는 방법 중 하나이다. 이산화 방법은 여러 가지 방법이 있을 수 있는데 본 연구에서는 다음의 3가지 방법을 고려한다.

(a) 각각의 변수의 값들이 독립적이라고 생각하고, 종속변수와 독립변수에 대한 고려 없이 이산화 시키는 방법(예 : equal frequency scaling, 전문가 지식을 이용하는 방법)
(b) 종속변수 값과의 관계를 고려하여 독립변수를 하나씩 이산화하는 방법(예 : minimum entropy scaling)
(c) 이산화 과정에서 종속변수의 값을 고려하여 모든 독립변수를 동시에 이산화 하는 방법(예 : naïve and Boolean reasoning-based discretization)

(a)의 경우에는 두 가지 세부 방법이 있으므로 크게 4가지의 방법들을 아래에서 설명하기로 한다.

2.3.1 등분위 이산화(equal frequency scaling)

이 방법은 변환이 가장 용이한 방법으로 연속형

자료를 각 독립변수 별 값의 크기순으로 정렬한 후, 이산화 된 후의 각각의 범주에 속한 개체의 개수가 거의 동일하도록 이산화를 해 주는 방법이다. 이 방법은 종속변수 값에 대한 고려가 없이 독립변수들을 이산화 시키는 방법이다.

2.3.2 전문가 지식에 의한 이산화

해당 분야의 전문가들에 대한 설문이나 인터뷰, 문헌분석 등에 의해서 경계값을 얻는 방법이다. 이 방법 역시 종속변수 값에 대한 고려가 없이 독립변수들을 이산화 시키는 방법이다.

2.3.3 최소 엔트로피 기준 이산화 (minimum entropy scaling)

이 방법은 Dougherty et al.(1995)에 의해 제안된 방법으로 이산화를 수행함에 있어서 각 범주의 엔트로피(entropy) 값이 최소가 될 수 있도록 하는 재귀 분할(recursive partitioning)을 이용하는 방법이다. 만일 표본집단 S가 $c_1, c_2, ..., c_k$의 집합으로 나누어지고, 각각이 $p_1, p_2, ..., p_k$의 확률값을 가진다면 집단 S의 엔트로피는 다음과 같은 식으로 계산할 수 있다.

$$Entropy(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

2.3.4 Naïve and Boolean reasoning 이산화

이는 Naïve scaling 후에 그로 인해 분할된 범주 값을 불리안 추론(Boolean reasoning) 방법으로 종속변수 값에 대해 최적화된 이산화 경계값을 구하는 방법이다(Nguyen and Skowron, 1995). Naïve scaling에 의한 이산화 방법은 다음과 같다. 모든 독립변수 값을 각각 크기순으로 정렬한다. 독립변수 중 한 변수에 대하여 다시 정렬한 후 그에 대하여 종속변수 값을 고려하여 종속변수 값이 변하는 부분의 두 개체의 독립변수 값을 평균한 것들을 한 독립변수의 경계값으로 삼는다. 이를 수식으로 나타내면 다음과 같다.

조건속성 $a(\in Q)$
속성 a에 대해 Sorting한 값:

$$v_a^1 < L < v_a^i < L < v_a^{cardinality(V_a)}$$

$$X_a^i = \{x \in U | a(x) = v_a^i\}$$

$$\Delta_a^i = \{v \in V_d \mid \exists\, x \in X_a^i \text{ such that } d(x) = v\}$$

$$C_a = \left\{ \frac{v_a^i + v_a^{i+1}}{2} \mid cardinality(\Delta_a^i) > \right.$$
$$\left. 1 \text{ or } cardinality(\Delta_a^{i+1}) > 1 \text{ or } \Delta_a^i \neq \Delta_a^{i+1} \right\}$$

$d(x)$: 의 의사결정속성값

이로 인해 얻어진 경계값을 그대로 수용한 후 전체 속성과의 연관 하에 조화를 이룰 수 있도록 하는 최소한의 경계값들의 집합을 구하기 위해 불리안 추론 방법을 이용한다. 불리안 POSITIVE 식은 "$h = \prod_{(x,y)} \sum_a \sum c^* \mid c \in C_a, a(x) < c < a(y), d_Q(x) = d_Q(y)$"로 사용한다.

위의 식을 만족하는 범위의 최소한의 경계값을 필요로 하는 경계값 집합을 구한다. 이에 대한 알고리즘은 Nguyen and Skowron(1995)에 나와 있다. 이 알고리즘은 2단계 이산화 방법론으로 1단계에서는 단일속성(하나의 독립변수)과 의사결정속성(종속변수)에서 구간경계 값을 구한 후, 2단계에서 전체 최적인 경계값 집합을 구하는 것이다.

3. 실험설계

본 연구에서는 전술한 바와 같이 주식시장 분석에 유용할 것으로 생각되는 러프집합분석을 이용

하여 선물지수에 대한 분석을 수행한다. 특히, 본 연구에서는 주식시장 분석에 러프집합분석을 적용할 때 반드시 필요한 최적 변수 이산화 방법을 살펴보고자 한다. 따라서 본 연구에서는 전술한 4가지 변수 이산화 방법에 따라 각 방법의 성과들을 비교해 볼 것이다.

러프집합분석을 실제 수행하는 과정에는 여러 가지 실험조건에 대한 고려가 선행되어야 한다. 실험조건에 따라 실험의 성과는 상이하게 나타날 수 있으므로 선행연구에서 일반적으로 사용되었던 여러 기준들을 기준으로 하여 본 연구의 실험을 수행한다.

본 연구에서는 러프집합분석을 수행한 후 얻어진 Reduct들을 이용하여 의사결정 규칙을 생성하고 이 중에서 지지도가 높은 규칙들을 우선적인 규칙집단으로 생성한다. 러프집합분석에서 규칙 내의 변수 선택 방법은 경제성에 기반한다. 즉, 전체 설명변수를 적용하였을 때 근사의 확실도가 거의 동일한 변수군 집합을 구한다. 설명변수가 많아질수록 규칙의 수가 많아지고 그에 비례하여 그 규칙이 적용될 수 있는 사례가 적어져서 각 규칙에 의해서 얻어지는 결과가 일반화되기 어렵다. 따라서 설명변수의 수를 줄이는 것이 결과의 일반화와 간결성을 위해서 중요하다.

러프집합분석을 통하여 일단의 Reduct를 구한 후 그 중 지지도가 가장 큰 것을 실험대상 속성집합으로 삼는다. 이는 주식시장 자료와 같이 잡음이 심한 자료의 경우에는 전체자료를 완전하게 설명할 수 있는 결정적인 규칙과 Reduct가 존재하기 어렵기 때문이다. 따라서 본 연구에서는 Reduct approximation이란 방법을 사용하는데, 이는 완전한 동질성을 보장할 수 있는 집합을 찾는 것이 아니라 동질적 관계를 거의 보장해 줄 수 있는 Reduct를 구하는 방법이다. 다음 단계로는 얻어진 Reduct들로부터 의사결정규칙을 생성해 낸다. 규칙을 생성하는 기준은 지지도가 20이 넘는 것과 규칙의 판별력이 50%를 넘는 것으로 한다. 지지도의 기준으로 20이란 값을 이용한 것은 본 연구에서 사용될 검증용 자료의 빈도수인 200여 개의 10% 수준을 의미하며, 이는 특정한 하나의 규칙이 최소한 전체 검증용 자료의 10% 이상을 설명해 주어야 한다는 것을 의미한다. 즉, 검증용 자료의 10% 수준을 설명할 수 있는 규칙들을 활용한다는 의미이다. 또한, 판별력의 값을 50%로 한 것은 최소한 50% 이상의 판별 확실성을 가져야 주식시장의 방향성을 파악하는데 유의하기 때문이다. 이러한 기준은 러프집합 자체에서는 자동으로 설정해 주지 못하기에 설계자의 주관적인 판단에 의해서 결정된다.

4. 거래 모의 실험

4.1 연구자료

주식시장에서 매매시점 결정은 초과수익을 얻기 위해 사용되는 투자전략의 하나이다. 전통적으로 초과수익은 주식시장에서의 주요한 전환시점을 기대하는 자산 구성에 의해 달성될 수 있는 것이다(Waksman et al., 1997). 본 연구에서의 매매시점 포착이란 거래로부터 얻을 수 있는 초과수익을 얻기 위해 매수 또는 매도의 시점을 결정하는 것을 의미한다. 일반적으로 예측기법이 매매시점 포착의 능력을 가지고 있는가를 판단하기 위해서는 통상 사용되던 통계적 검증방법과는 다른 방법이 요구되는데 Merton(1981)은 예측기법이 매매시점 포착의 능력을 갖기 위한 이론적 근거를 제시하였고, Henriksson and Merton(1981)은 이 이론에 대한 검증방법을 제시한 바 있다. 그러나 이러한 통계적인 검증방법을 통하지 않더라도 매수

후보유전략(buy and hold strategy)에 의한 기간 수익률과 시스템에서 산출되는 기간수익률을 비교함으로써 용이하게 시스템의 매매시점 포착능력을 확인할 수 있다.

본 연구에서는 실제 주가지수 선물시장의 자료를 이용하여 러프집합이론의 적용 가능성을 검증해 보고자 한다. 분석을 위해 사용된 데이터는 한국 주가지수선물(KOSPI 200 stock index futures) 가격 자료이다. 이 시장은 1996년 5월 3일에 개장하였는데, 사용된 자료의 기간은 1996년 5월 14일부터 1998년 10월 14일까지의 660거래일의 일별 자료이다. 그 중 1998년 자료는 시스템의 일반화 능력 검증을 위해 사용한다. 실험을 위한 자료의 구성내용은 <표 1>과 같다.

<표 1> 실험 자료 구성

	학습용 데이터	검증용 데이터
기간	1996년 5월 ~ 1997년 11월(428일)	1998년 1월 ~ 1998년 10월(282일)

선물가격자료는 일반적으로 동일시점에서 4개의 만기일 별로 4개의 가격이 존재한다. 따라서 어떤 가격자료를 이용하느냐에 따라서 자료의 구성이 달라지게 된다. 본 연구에서는 최근 월물자료 연속사용법(장재건 등, 1996)을 이용하여 선물가격자료를 수집하였다. 이 방법은 가장 최근에 도래하는 만기일에 해당하는 선물가격자료를 이용하는 방법으로 일반적으로 선물시장분석에 많이 이용되는 방법이다.

본 연구에서 얻고자 하는 결과물은 매일의 매매시점 여부이므로 단기예측에 해당한다고 할 수 있다. 따라서 본 연구에서는 전술한 선물가격자료를 이용하여 단기분석에 유용한 것으로 알려진 기술적 지표를 생성시켰다. 러프집합에 관한 이론연구에서 언급한 바와 같이 러프집합을 이용하면 유의한 변수군을 러프집합이 선정해 주므로 본 연구에서는 기존 연구에서 많이 사용되고 있는 9개의 기술적 지표를 기본 변수군으로 사용하였다. 각 기술적 지표의 명칭과 기술통계량은 <표 2>에 제시되어 있다.

러프집합분석은 <표 2>에서 제시된 기술적 지표 중에서 분석에 유용한 변수군을 다시 선정하는 과정을 거치게 된다. 전술한 바와 같이 러프집합분석은 변수 이산화 과정을 거치고 난 후에 분석을 수행하게 되므로 먼저 변수 이산화 과정을 언급하여야 하나, 이는 본 연구에서 가장 관심을 갖고 연구할 내용이기에 다음 절에서 상술하도록 하도록 한다.

<표 2> 기본 변수 군과 기술통계량

변수기호(변수명)	최대값	최소값	평 균	표준 편차
A1(Stochastic %K)	151.02	0.00	43.69	33.77
A2(Stochastic %D)	118.57	0.00	43.65	28.70
A3(RSI)	100.00	0.00	43.58	29.21
A4(Momemtum)	10.90	-11.35	-0.44	3.24
A5(ROC)	129.14	81.51	99.55	5.86
A6(A/D Oscillator)	1.71	-0.10	0.48	0.32
A7(CCI)	229.14	-212.14	-12.89	81.59
A8(Price oscillator)	8.14	-9.00	-0.39	9.72
A9(Disparity 5days)	114.72	87.25	99.72	3.16

4.2 러프집합분석과 이산화 방법

전술한 바와 같이 러프집합분석을 위해서는 실험에 사용할 기술적인 지표들을 이산화하여야 한다. 본 연구에서는 선행연구에서 제시한 4가지의 기존 이산화 방법을 비교 분석하도록 한다. 이산화를 위해 사용한 4가지 방법으로는 비지도적 방법(unsupervised method)인 등분위 이산화 방법과 전문가 지식 기반 이산화 방법이 있으며, 지도적 방법(supervised method)인 최소 엔트로피 기준 이산화 방법과 Naïve and Boolean reasoning 이산화 방법으로 나눌 수 있다. 이 중 전문가 지식에 의한 방법이 기존의 러프집합분석에서 주로 이용하던 방법이다. 아래에서는 각각의 이산화 방법을 본 연구의 실험데이터에 적용하여 러프집합분석을 수행했을 때의 성과를 정리한다.

4.2.1 등분위 이산화 방법을 이용한 분석결과

등분위 이산화 방법에서는 각각의 기술적인 지표들을 3개의 동일한 원소수의 집합으로 만든다. 집단을 3분위수를 기준으로 나누어 한 변수 당 3개의 값을 가질 수 있도록 변수를 변환한다. 그 결과 나타난 각 변수의 범주값의 범위는 <표 3>과 같다.

<표 3> 등분위 이산화 결과

변 수	범주		
	1	2	3
A1(Stochastic %K)	~18.0526	~58.4524	58.4524~
A2(Stochastic %D)	~20.2209	~57.1205	57.1205~
A3(RSI)	~22.7553	~56.5942	56.5942~
A4(Momemtum)	~-2.0250	~0.5750	0.5750~
A5(ROC)	~97.5328	~100.7610	100.7610~
A6(A/D Oscillator)	~0.2450	~0.6508	0.6508~
A7(CCI)	~-60.6931	~12.5319	12.5319~
A8(OSCP)	~-1.1254	~0.1920	0.1920~
A9(Disparity 5days)	~98.8529	~100.2320	100.2320~

이를 통해 만들어진 이산형 자료를 가지고 418개의 Reduct를 구하였고, 이 중 최대 지지도를 가지는 Reduct 규칙 군의 규칙들은 <표 4>와 같다.

<표 4> 등분위 이산화에 의해 도출된 거래 규칙

규 칙	조건			의사결정	
	A5	A6	A8	예측	Strength
#1	1	1	1	UP	45
#2	3	3	3	UP	43
#3	3	2	3	UP	36
#4	1	2	1	DOWN	25
#5	2	2	2	DOWN	24
#6	2	1	2	DOWN	23
#7	1	1	2	DOWN	22

한편, 러프집합분석에 의해 생성된 규칙을 평가하기 위해서는 시장에서 실제 거래에 이용하기 위한 거래 전략에 적용하여야 한다. 러프집합분석을 통해 생성되는 의사결정규칙은 “강세장”과 “약세장”의 두 가지 신호를 출력하거나 패턴이 불확실한 개체(거래일)에 대해서는 아무런 신호를 출력하지 않지만, 투자자는 이미 매수하고 있거나 매수하지 않은 상태인 두 가지의 상황이 존재할 수 있으므로 각각의 상황과 시스템에서 출력되는 신호를 실제 거래에 이용하기 위해서는 합리적인 거래 전략을 설정하여야 한다. 본 연구에서 사용된 거래 전략은 <그림 2>와 같다.

```
IF 당일의 시스템 생성신호 = “강세장”(익일기준)
    And if 전일의 의사결정 = “매수”;
        Then “매도”;
        Else “매수”;
IF 당일의 시스템 생성신호 = “ ”(대응하는 규칙이 존재하
                              지 않는 경우)
    Or if 당일의 시스템 생성신호 = “약세장”(익일기준)
        And if 전일의 의사결정 = “매수”
        Then “매도”;
    Else 투자 보류
```

<그림 2> 거래 전략

<표 4>에서 규칙 #1의 경우, 변수 A5, A6, A8이 각각 “1”의 값을 가질 때 적용되는 규칙이며, 그 결과값은 생성신호로 나타나는데 익일 기준으로 “강세장(UP)”임을 나타내고, 이 규칙을 적용받는 학습용 자료의 개수가 45개임을 나타낸다. <표 4>에 제시된 규칙들을 <그림 3>의 거래전략에 적용하여 학습용 자료와 검증용 자료에 적용한 결과와 “매수 후 보유 (Buy and hold)” 전략에 의한 결과가 <표 5>에 제시되어 있다.

<표 5> 등분위 이산화에 의한 거래 결과

Reduct	학습용 데이터			검증용 데이터		
	잔액	수익률	샤프지수	잔액	수익률	샤프지수
{A5, A6, A8}	12,035	20.35%	0.0556	11,725	17.25%	0.0462
Buy and Hold	4,373	-56.27%	-0.0975	9,663	-3.37%	0.0164

4.2.2 전문가 지식 기반의 이산화 방법의 분석결과

전문가 지식에 의한 이산화 방법은 전술한 바와 같이 해당 분야 전문가의 지식을 활용하여 이산화의 기준으로 삼는 것이다. 주식시장 전문가의 지식을 얻는 방법은 주식거래를 직접 하는 투자전문가의 의견을 이용할 수도 있고, 주식시장 분석에 관련된 선행연구들의 결과를 활용하는 방법도 있다. 본 연구에서 사용된 기술적 지표들은 경험적 연구의 결과로 이미 선행연구에 전문가의 해석기준이 충분히 제시되어 있으므로 본 연구에서는 선행연구의 내용을 이용하여 전문가의 지식을 얻도록 한다. 본 연구에서는 선행 연구 중 Murphy(1986), Achelis(1995), Gifford(1995), 장재건 등(1996), Edwards and Magee(1997) 등을 활용하여 이산화의 경계값을 생성하였다. 이 방법에 의해 얻어진 경계값의 범위는 <표 6>에 정리된 것과 같다.

<표 6> 전문가 지식에 의한 이산화 결과

변 수	범주		
	1	2	3
A1(Stochastic %K)	~25	~75	75~
A2(Stochastic %D)	~25	~75	75~
A3(RSI)	~30	~70	70~
A4(Momemtum)	~0	0~	
A5(ROC)	~100	100~	
A6(A/D Oscillator)	~0.5	0.5~	
A7(CCI)	~0	0~	
A8(OSCP)	~0	0~	
A9(Disparity 5days)	~100	100~	

이 과정을 통해 이산화 된 자료를 이용하여 280개의 Reduct를 구하였고, 이 중 지지도가 20이상인 것들 중 판별력이 50% 이상인 규칙을 가진 규칙 군만을 선별하여 규칙을 생성한다. 이에 따라 생성된 의사 결정 규칙은 <표 7>과 같다.

<표 7> 전문가 지식 기반 이산화에 의한 거래 규칙

규 칙	조건				의사 결정	
	A1	A3	A6	A8	예측	Strength
#1	2	2	2	1	Up	29
#2	2	2	1	1	Down	21
#3	3	3	2	2	Up	47
#4	2	2	1	2	Down	21
#5	1	1	2	1	Down	21
#6	2	1	2	1	Down	33
#7	1	1	1	1	Down	104

한편, 생성된 규칙에 의한 투자성과는 <표 8>과 같다. <표 8>에서 나와 있는 것처럼 가장 성과가 좋은 규칙 군은{A1, A3, A6, A8}으로 “Buy and Hold(매수 후 보유)” 전략보다 좋은 성과를 보여준다.

<표 8> 전문가 지식 기반 이산화에 의한 거래 결과

Reduct	학습용 데이터			검증용 데이터		
	잔액	수익률	샤프지수	잔액	수익률	샤프지수
{A1, A3, A6, A8}	10,830	8.30%	0.0310	13,727	37.27%	0.0868
Buy and Hold	4,373	-56.27%	-0.0975	9,663	-3.37%	0.0164

4.2.3 최소 엔트로피 기준 이산화 방법의 분석결과

이 방법에 의해서는 본 연구의 자료에 대한 이산화가 거의 이루어지지 않았다. 이는 주식시장 자료 자체가 잡음이 너무 심하여 최소 엔트로피에 의한 경계값을 찾기가 어려운 이유 때문인 것으로 판단된다. 분석의 결과와 그 성과는 <표 9>와 같다.

<표 9> 최소 엔트로피 기준 이산화에 의한 거래 결과

Reduct	학습용 데이터			검증용 데이터		
	잔액	수익률	샤프지수	잔액	수익률	샤프지수
{A3}	9,833	-1.67%	-0.0093	9,857	-1.43%	-0.0658
Buy and Hold	4,373	-56.27%	-0.0975	9,663	-3.37%	0.0164

<표 9>에서 보는 바와 같이 {A3}만이 유의한 것으로 생성되었으며, 이를 통한 거래로는 양의 수익을 얻을 수 없었다. 기존의 연구에서는 최소 엔트로피 기준의 이산화 방법이 가장 우수한 것으로 보고되었으나, 본 연구에서는 좋은 성과를 얻을 수 없었으며, 이는 잡음이 심한 주식시장 데이터의 특성 때문인 것으로 생각된다.

4.2.4 Naïve and Boolean reasoning 이산화 방법의 분석결과

이 방법은 두 단계의 이산화 과정을 거친다. 즉, 초기 이산화 후 이를 다시 전체적으로 최적화(경계점 집합의 수를 최소화)하는 형태로 이루어진다. 이 방법에 의한 경계값의 범위는 <표 10>과 같다.

<표 10>의 기준에 의한 이산화 후, 러프집합분석에 의해 얻어진 규칙들은 <표 11>과 같다.

<표 11> Naïve and Boolean Reasoning 이산화에 의한 거래 규칙

규칙	조건			의사결정	
	A4	A6	A9	예측	Strength
#1	2	1	1	DOWN	37
#2	1	2	1	DOWN	22
#3	1	1	1	DOWN	35
#4	2	3	1	UP	32
#5	2	2	1	DOWN	34
#6	4	3	1	UP	23
#7	4	2	1	UP	31

<표 10> Naïve and Boolean Reasoning에 의한 이산화 결과

변 수	범주						
	1	2	3	4	5	6	7
A1(Stochastic %K)	~2.5645	~22.6111	~58.6336	~69.0933	~91.9373	~96.7342	96.7342~
A2(Stochastic %D)	~8.5176	~19.1029	~33.3168	~49.4679	~85.9901	85.9901~	
A3(RSI)	~7.1180	7.1180~					
A4(Momemtum)	~-2.9250	~-0.9750	~0.0000	~2.6500	2.6500~		
A5(ROC)	~95.3582	~96.4081	96.4081~				
A6(A/D Oscillator)	~0.1699	~0.4298	~0.7244	~0.9075	0.9075~		
A7(CCI)	~-52.6728	~-26.2145	~50.1580	~93.5171	93.5171~		
A8(OSCP)	~-1.4078	~-0.5346	~0.1229	~0.6543	0.6543~		
A9(Disparity 5days)	~102.0040	102.0040~					

한편, <표 11>에 나타난 규칙 군을 이용한 모의 거래 결과는 <표 12>와 같다.

<표 12> Naïve and Boolean Reasoning 이산화에 의한 거래 결과

Reduct	학습용 데이터			검증용 데이터		
	잔액	수익률	샤프지수	잔액	수익률	샤프지수
{A4, A6, A9}	13,615	36.15%	0.0956	9,483	-5.17%	-0.0084
Buy and Hold	4,373	-56.27%	-0.0975	9,663	-3.37%	0.0164

<표 12>의 결과를 보면, 이 방법의 경우에는 학습용 자료에서는 이익이 발생하였으나, 검증용 자료에서는 이익이 발생하지 않았음을 알 수 있다.

5. 실험결과의 종합분석

5.1 러프집합분석의 효과와 변수 이산화 방법 간의 종합비교

실험 결과, 일반적인 러프집합분석에 의해 생성한 규칙에 의한 거래가 'Buy and Hold' 전략보다는 우월한 결과를 보였다. 그러나 이산화 방법에 따라서는 상당히 많은 차이점을 보였다. 학습용 자료에서 가장 높은 수익률을 보이는 규칙 군을 각각의 이산화 방법들에서 하나씩 선택하여 그들을 비교한 결과는 <표 13>과 같다.

또한 각 이산화 방법의 학습용 자료와 검증용 자료에서의 성과 추이를 나타낸 것이 <그림 3>이다.

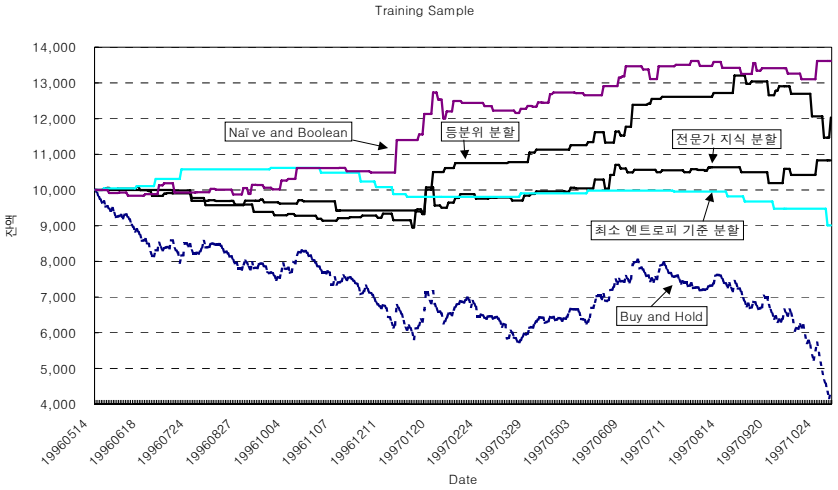
<표 13>과 <그림 3>을 통해서 검증용 자료에서는 전문가 지식에 의한 변수 이산화 방법의 성과가 가장 좋음을 알 수 있다. Naïve and Boolean Reasoning에 의한 이산화 방법의 경우에는 학습용 자료에서는 성과가 좋으나, 검증용 자료에서는 성과가 좋지 않아서 일반화된 성과를 보여 주지 못하였다. 최소 엔트로피 기준 이산화 방법의 경우에는 생성된 결과로 보아 잡음이 심한 주식시장 자료에서는 이산화가 효과적으로 이루어지지 못한 것으로 판단된다.

이상의 결과를 통하여, 학습용과 검증용 자료에서의 성과의 유사성을 기준으로 안정성을 판단할 때, 안정성을 가진 이산화 방법으로 생각되는 것은 등분위 분할 이산화 방법과 전문가 지식에 의한 이산화 방법으로 판단된다.

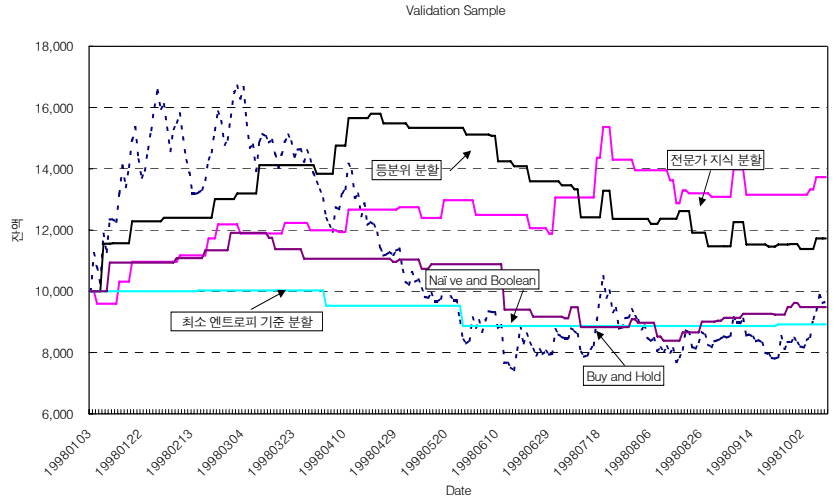
따라서, 학습용 자료와 검증용 자료에서의 성과와 안정성 등을 감안하여 생각해 볼 때, 기존의 러프집합분석에서 이산화 방법으로 주로 쓰이던 전문가 지식에 의한 분할의 방법이 가장 우수한 것으로 나타났다. 그러나, 이 방법을 위해서는 전문

<표 13> 이산화 방법에 따른 성과 비교

이산화 방법	학습용 데이터			검증용 데이터		
	잔액	수익률	샤프 지수	잔액	수익률	샤프 지수
전문가지식	10,830	8.30%	0.0310	13,727	37.27%	0.0868
등분위 분할	12,035	20.35%	0.0556	11,725	17.25%	0.0462
최소 엔트로피 기준	9,833	-1.67%	-0.0093	9,857	-1.43%	-0.0658
Naïve and Boolean Reasoning	13,615	36.15%	0.0956	9,483	-5.17%	-0.0084
Buy and Hold	4,373	-56.27%	-0.0975	9,663	-3.37%	0.0164



(a) 학습용 데이터



(b) 검증용 데이터

<그림 3> 이산화 방법들의 성과비교

가의 지식을 습득해야 하고, 전문가의 지식을 추출하기 어려운 분야에서는 사용하기 어렵다는 문제점이 존재한다. 이에 대한 차선의 방법으로는 안정적이고 성과가 좋은 등분위 분할 이산화 방법을 생각할 수 있는데, 이 또한 처리과정의 논리적인 근거가 부족하다는 점에서 문제가 있다.

5.2 러프집합분석과 다른 규칙 생성 방법과의 비교

추가적인 검증단계로 기존 규칙 생성 방법 중 일반적으로 가장 많이 쓰이는 의사결정나무분석과 본 연구의 결과를 비교해 보았다. 본 분석에서

<표 14> 러프집합분석과 의사결정나무분석(C4.5)을 이용한 거래결과 비교

	학습용 데이터			검증용 데이터		
	잔액	수익률	샤프지수	잔액	수익률	샤프지수
전문가 지식 분할 후 러프 집합 분석	10,830	8.30%	0.0310	13,727	37.27%	0.0868
전문가 지식 분할 후 C 4.5분석	9430	-5.70%	-0.0274	13074	30.74%	0.1193
Buy and Hold	4,373	-56.27%	-0.0975	9,663	-3.37%	0.0164

는 러프집합분석에서 가장 좋은 성과를 보인 전문가 지식을 이용한 이산화 방법의 범주 범위를 의사결정나무분석의 하나인 C4.5(Quinlan, 1993)에 적용하여 규칙을 생성하였다. 러프집합분석에 의해 도출된 규칙에 의한 거래와 의사결정나무분석에 의해 도출된 규칙을 이용한 거래를 모의 실험해 본 결과는 <표 14>와 같다.

<표 14>를 통해 검증용 자료에서 러프집합분석의 성과가 C4.5의 성과보다 좀 더 우수한 것을 확인할 수 있다. 따라서 본 연구에서 제안하는 방법이 기존의 규칙 생성 방법에 비해서도 성과가 나쁘지 않음을 확인할 수 있다.

6. 결론

본 연구에서는 한국 주가지수 선물시장에서의 매매시점을 결정하기 위해 러프집합분석을 이용하였고, 러프집합분석에서 사용되는 변수 이산화 방법들의 투자성과차이를 확인하였다. 러프집합분석의 성과는 대체로 "매수 후 보유" 전략의 성과보다 우수한 것으로 나타났으며, 이는 러프집합분석이 자료 내에 포함되어 있는 잡음과 불확실한 패턴을 제거하고, 확실한 패턴에 대해 과거 자료에 기반한 결과를 제시해 주기 때문인 것으로 판단된다. 이는 러프집합분석이 지지도와 판별력 등의 자체 통제기능을 이용하여 유의한 패턴을 찾아 주기에 가능한 것이라고 생각된다. 이는 C4.5의 성과비교에서도 확인할 수 있었다.

본 연구는 러프집합분석의 주식시장 분석 가능성 확인 이외에, 러프집합분석에서 반드시 선행되어야 하는 연속형 자료의 이산화하는데 사용되는 여러 가지 이산화 방법들의 성과를 비교 분석하였다. 본 연구의 분석 결과, 일반적으로 이용되는 4가지 이산화 방법 중 본 연구의 자료에서는 전문가 지식을 이용한 이산화 방법의 성과가 가장 우수하고 성과의 안정성도 높은 것으로 나타났다. 다른 방법들은 전문가 지식을 이용한 이산화 방법보다 일반화의 정도도 낮고, 성과가 좋지 못한 것으로 나타났다. 따라서 본 연구의 결과에 의하면, 주식시장의 매매시점을 위한 러프집합분석에서는 전문가 지식을 활용한 이산화 방법을 활용하는 것이 유용할 것으로 판단된다. 한편, 전문가 지식이 존재하지 않거나 쉽게 추출할 수 없는 분야에서는 등분위 분할 이산화 방법도 사용될 수 있을 것으로 생각된다. 이 방법은 성과 면에서 전문가 지식을 이용한 방법보다 좋지 못하였으나 일반화의 가능성은 높은 것으로 나타났다.

본 연구는 이상과 같은 연구결과를 제시하였지만 많은 한계점도 가지고 있다. 첫째, 본 연구에서는 거래에 따른 거래비용을 고려하지 않았다. 그러나, 러프집합분석은 모든 거래일에 대해 매매신호를 생성하지 않고 확실한 패턴을 가진 거래일에만 매매신호를 생성해 주므로 다른 방법에 비해 거래비용의 효과는 크지 않을 것으로 예상된다. 둘째,

본 연구에서 사용된 실험자료의 수가 결과를 일반화하기에 충분치 않다. 이는 러프집합분석이 실험자료가 많을 때에는 분석시간이 오래 걸리고 분석도 복잡해지는 현실적인 한계에 기인한 것이나 향후 연구에서는 보다 충분한 실험자료의 확보를 통해 더 일반화된 연구결과를 제시할 수 있을 것이다. 셋째, 투자성과를 계산하기 위한 거래 전략이 주관적이다. 연구과정의 단순화를 위하여 거래전략을 구성하였으나 향후 연구에서는 본 연구에서 제시된 거래 전략 이외에도 다양한 전략에 의한 결과가 비교되어야 할 것이다. 넷째, 다양한 연구 데이터에 대한 실험을 통한 일반화된 결과를 도출하지 못하였다. 본 연구의 주제가 주식시장의 매매시점을 결정하기 위한 분석이므로, 주식시장의 데이터를 활용하였으나, 보다 다양한 재무 시계열 데이터를 활용하여 분석하였다면 연구 결과의 일반화를 할 수 있었을 것이다.

참고문헌

김창연, 안병석, 조성식, 김성희, 도산예측을 위한 러프집합 이론과 인공신경망 통합방법론, *경영정보학연구*, 9권 4호(1999), 23~40.

박기남, 이훈영, 박상국, 러프집합을 이용한 통합형 채권등급 평가모형 구축에 관한 연구, *한국경영과학회지*, 20권 3호(2000), 125~135.

장재건, 정용만, 연광제, 전준우, 신동현, 김현태, 기술적 분석지표를 이용한 선물투자기법, 도서출판 *진리탐구*, 1996.

정석훈, 서용무, Rough Set 기법을 이용한 신용카드 연체자 분류, Entrue Journal of Information Systems, 7권 1호(2008), 141~150.

Achelis, S. B., *Technical analysis from A to Z*. Chicago : Probus Publishing, (1995).

Ahn, B. S., Cho. S. S. and Kim. C. Y., The integrated methodology of rough set theory and artificial neural network for business failure prediction., *Expert Systems with Applications*, Vol.18, No.2(2000), 65~74.

Dimitras, A. I., Slowinski. R., Susmaga. R. and Zopounidis. C., Business failure prediction using rough sets, *European Journal of Operational Rese*arch, Vol.114, No.2(1999), 263~280.

Edwards, R. D. and Magee. J., *Technical analysis of stock trends*, Chicago, Illinois : John Magee (1997).

Grzymala-Busse, J. W., LERS-s system for learning from examples based on rough sets. In R. Slowinski (Ed.), *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*(3~18), Kluwer Academic Publisher(1992).

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, No.2(1998), 390~402.

Henriksson, R. D. and Merton. R. C. On market timing and investment performance II : Statistical procedures for evaluating forecasting skill, *Journal of Business*, Vol.54(1981), 513~533.

Kim, K. and Han. I. The extraction of trading rules from stock market data using rough sets, *Expert Systems*, Vol.18, No.4(2001), 194~202.

Merton, R. C. On market timing and investment performance I : An equilibrium theory of value for market forecasts. *Journal of Business*, Vol.54(1981), 363~406.

Murphy, J. J., *Technical analysis of the futures*

markets : A comprehensive guide to trading methods and applications, New York : Prentice-Hall(1986).

Nguyen, H. S. and Skowron. A. Quantization of real-valued attributes, In *Proc. Second International Joint Conference on Information Sciences*, Wrightsville Beach, NC, (1995), 34~37.

Pawlak, Z., Rough sets, *International Journal of Information and Computer Sciences*, Vol.11 (1982), 341~356.

Pawlak, Z., Rough set approach to knowledge -based decision support. *European Journal of Operational Research*, Vol.99(1997), 48~57.

Quinlan, J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, (1993).

Ruggiero, M. A., *Cybernetics trading strategies : Developing profitable trading systems with state-of-the-art technologies*, New York : John Wiley and Sons, (1997).

Skowron, A., Boolean reasoning for decision rules generation. In J. Komorowski. and Z. W. Ras. (Eds.), *Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence*(1993), 295~305.

Slowinski, R. and Stefanowski. J., Rough classification with valued closeness relation. In E. Diaday et al. (Eds.), *New Approaches in Classification and Data Analysis*(1994), 482~488.

Slowinski, R. and Zopounidis. C., Application of the rough set approach to evaluation of bankruptcy risk, *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.4, No.1(1995), 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, No.1(1997), 1~15.

Tay, F. E. H. and Shen. L., Economic and financial prediction using rough sets model, *European Journal of Operational Research*, Vol.141, No.3(2002). 641~659.

Waksman, G., Sandler. M., Ward. M. and Firer. C., Market timing on the Johannesburg stock exchange using derivative instruments, *Omega*, Vol.25, No.1(1997), 81~91.

Yeh, C. C., Chi. D. J. and Hsu. M. F., A hybrid approach of DEA, rough set and support vector machines for business failure prediction, *Expert Systems with Applications*, Vol. 37(2010), 1535~1541.

Abstract

Rough Set Analysis for Stock Market Timing

Jin-nyung Huh · Kyoung-jae Kim · Ingoo Han

Market timing is an investment strategy which is used for obtaining excessive return from financial market. In general, detection ofmarket timing means determining when to buy and sell to get excess return from trading. In many market timing systems, trading rules have been used as an engine to generate signals for trade. On the other hand, some researchers proposed the rough set analysis as a proper tool for market timing because it does not generate a signal for trade when the pattern of the market is uncertain by using the control function. The data for the rough set analysis should be discretized of numeric value because the rough set only accepts categorical data for analysis. Discretization searches for proper “cuts” for numeric data that determine intervals. All values that lie within each interval are transformed into same value. In general, there are four methods for data discretization in rough set analysis including equal frequency scaling, expert’s knowledge-based discretization, minimum entropy scaling, and naïve and Boolean reasoning-based discretization. Equal frequency scaling fixes a number of intervals and examines the histogram of each variable, then determines cuts so that approximately the same number of samples fall into each of the intervals. Expert’s knowledge-based discretization determines cuts according to knowledgeof domain experts through literature review or interview with experts. Minimum entropy scaling implements the algorithm based on recursively partitioning the value set of each variable so that a local measure of entropy is optimized. Naïve and Booleanreasoning-based discretization searches categorical values by using Naïve scaling the data, then finds the optimized dicretization thresholds through Boolean reasoning. Although the rough set analysis is promising for market timing, there is littleresearch on the impact of the various data discretization methods on performance from trading using the rough set analysis. In this study, we compare stock market timing models using rough set analysis with various data discretization methods. The research data used in this study are the KOSPI 200 from May 1996 to October 1998. KOSPI 200 is the underlying index of the KOSPI 200 futures which is the first derivative instrument in the Korean stock market. The KOSPI 200 is a market value weighted index which consists of 200 stocks selected by criteria on liquidity and their status in corresponding industry including manufacturing, construction, communication, electricity and gas, distribution and services, and financing. The total number of samples is 660 trading days. In addition, this study uses popular

technical indicators as independent variables. The experimental results show that the most profitable method for the training sample is the naïve and Boolean reasoning but the expert's knowledge-based discretization is the most profitable method for the validation sample. In addition, the expert's knowledge-based discretization produced robust performance for both of training and validation sample. We also compared rough set analysis and decision tree. This study experimented C4.5 for the comparison purpose. The results show that rough set analysis with expert's knowledge-based discretization produced more profitable rules than C4.5.

Key Words : Rough Set, Market Timing, Discretization, Expert's Knowledge, Profitability

저 자 소 개

허진영

서울대학교 경영학과 학사, KAIST 경영공학 석사를 졸업한 후 한국오라클에서 일한 후 현재는 ㈜네오아이즈 이사로 재직 중이다. 주요 수행 업무는 금융기관 위험관리 및 관리회계 부분이며, 현재는 금융기관의 국제회계기준(IFRS) 시스템 프로젝트에서 공정가치, 수익인식, 대손충당금과 관련된 부분의 컨설팅을 하고 있다.

김경재

현재 동국대학교 경영대학 경영정보학과 부교수로 재직 중이다. KAIST에서 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 신용평가 시스템, 지식경영, 고객관계관리 등이다.

한인구

서울대학교 국제경제학 학사, KAIST 경영과학 석사, University of Illinois at Urbana-Champaign에서 회계정보시스템을 전공하여 경영학박사를 취득하고 KAIST 경영대학 교수로 재직 중이다. 주요 연구분야는 회계 및 재무 분야에서의 인공지능 응용, 지식자산, 온라인 고객행동분석 등이다.