한국지능정보시스템학회논문지 제12권 제4호 2006년 12월(pp. 15~37)

데이터마이닝기법을 이용한 주식시장의 이상매매 적출

홍정훈
국민대학교 경영대학
(chhong@kookmin.ac.kr)

안성만
국민대학교 경영대학
(sahn@kookmin.ac.kr)

위경우
숙명여자대학교 경영학부
(kwwee@sookmyung.ac.kr)

본 논문은 증권거래소 이상매매 적출업무의 효율성을 제고하기 위해 데이터마이닝 기법을 적용하는 방안에 대해 연구하는 것을 주된 목적으로 한다. 이 과정에서 국내 증권거래소의 이상매매 적출모형과 데이터마이닝을 활용한 해외사례로서 미국 NASD의 ADS를 소개한 뒤, 실증분석에 사용될 자료들을 시세조종 종목과 정상 종목으로 나누어 검토한다. 국내에서 주식시장의 이상매매 적출에 대한 데이터마이닝 기법의 적용에 대한 연구가 없는 상황에서 다양한 입력변수를 만들어 실제로 데이터마이닝 기법들을 적용하여 적출성과를 상호 비교한 결과와 시사점을 기술하였다.

논문접수일 : 2006년 03월 게재확정일 : 2006년 12월 교신저자 : 안성만

1. 서론

본 논문은 증권거래소 이상매매 적출업무의 효율성을 제고하기 위해 데이터마이닝 기법을 적용하는 방안에 대해 조사·분석하는 것을 주된 목적으로 한다. 현재 증권거래소에서는 주식시장의 이상매매 적출에 대한 데이터마이닝 기법을 적용하고 있지 않으며, 이에 대한 연구도 없는 상황이므로 본 논문이 그 분야에서의 데이터마이닝 기법의 적용에 대한 가능성을 열어 놓기를 기대한다.

본 논문은 우선 현재 증권거래소에서 사용하고 있는 이상매매 적출모형을 개괄하고, 동 업무에 데이터마이닝 기법을 활용하고 있는 대표적인 사례로서 미국 NASD (National Association of Securities Dealers)의 ADS (Advanced Detection System)를 소개한다.

다음으로 본 논문에서는 거래소에서 축적해 온 시세조종 종목과 정상 종목들의 과거 자료들을 기초로 이상매매 적출모형을 구축하기 위한 최적 데이터마이닝 기법을 선정하고, 적출모형에 필요한 구체적인 변수들을 선정하는 실증분석을 실시한다. 이상매매의 적출에 사용될 데이터마이닝 기법으로는 로짓모형, 신경망, 의사결정나무 등 지도학습에 속하는 기법들이 선정과정에서 후보가 될 수 있으며, 변수들 사이의 규칙을 발견하기 위한 연관분석도 고려 가능한 것으로 판단된다. 검토대상 변수로는 기업의 일일거래 자료와 일중거래 자료들을 모두 포함하여 분석을 시도한다.

실증분석의 결과 국내의 경우 이상매매 적출을 위한 데이터마이닝 기법으로는 의사결정나무와 로짓 모형을 활용하는 것이 가장 바람직한 것으로 나타났다.

본 논문은 다음과 같이 구성된다. 2절은 국내 증권거래소의 이상매매 적출모형과 데이터마이닝을

활용한 해외사례로서 미국 NASD의 ADS를 소개한다. 3절은 실증분석에 사용될 자료들을 시세조종 종목과 정상 종목으로 나누어 검토한다. 4절은 실제로 데이터마이닝 기법들을 적용하여 적출성과를 상호 비교하고, 마지막으로 5절에서는 본 연구의 시사점과 제안사항들을 제시한 후 마무리한다.

2. 국내외 이상매매 적출모형의 현황 및 사례

2.1 증권거래소의 이상매매 적출모형

현재 증권거래소는 심리업무의 출발점이 되는 이상매매 적출모형으로서 통계적 회귀분석 방법에 기초한 '시장주가감시' 모형과 '기간주가감시' 모형을 사용하고 있다. 이들에 대해 간략히 살펴보면 다음과 같다.

2.1.1 시장주가감시 모형

시장주가감시 모형은 일정한 시점에서 주어진 유의수준 하에서 특정주식의 주가수익률이 그 시점에서의 정상적인 변동범위를 벗어나거나, 거래량의 회전율이 그 시점에서의 정상적인 변동범위를 벗어나는 상황을 찾아내는 방법이다.

시장주가감시 모형은 주가기준 모형과 거래량기준 모형으로 구성되어 있고, 각각의 모형은 평균조정모형, 2요인 시장모형, 1요인 시장모형으로 이루어져 있다. 여기서 평균조정모형은 특정주식에 대한 평균수익률과 평균거래회전율을 기준으로 한 모형이고, 2요인 시장모형은 특정 주식의 수익률(거래회전율)을 종합주가지수 수익률(시장가중평균 거래량)과 업종지수 수익률(업종가중평균 거래량)의 2가지 수익률(거래량) 대해 회귀분석을 수행하여 도출된 예상수익률(예상 거래회전율)을 기준으로 한 모형이다. 한편 1요인 시장모형은 특정 주식의 수익률(거래회전율)을 종합주가지수 수익률(시장가중평균 거래량)과 업종지수 수익률(업종가중평균 거래량)에 대해 각각 회귀분석을 수행하여 그 가운데 R^2값이 큰 모형을 선택하여 도출된 예상수익률(예상 거래회전율)을 기준으로 한 모형이다.

현재 증권거래소 적용하고 있는 시장주가감시 모형은 주가기준 모형과 거래량기준 모형 중 평균조정모형(유의수준 5%와 4%)과 2요인 시장모형(유의수준 1%와 2%)인데, 적용모형과 유의수준을 시황상황에 따라 탄력적으로 변경하여 사용하고 있다.

2.1.2 기간주가감시 모형

기간주가감시 모형은 일정한 시점에서 주어진 유의수준 하에서 특정기업의 누적 비정상수익률이 그 시점에서의 정상적인 변동범위를 벗어나거나, 누적 비정상거래량의 회전율이 그 시점에서의 정상적인 변동범위를 벗어나는 상황을 찾아내는 방법이다.

기간주가감시 모형은 주가기준 모형과 거래량기준 모형, 그리고 관여도 기준모형의 세 부분으로 구성되어 있고, 주가기준 모형과 거래량기준 모형은 평균조정모형, 2요인 시장모형, 1요인 시장모형으로 이루어져 있다. 여기서 관여도는 지점관여도(BCR : Branch Concentration Ratio)와 계좌관여도(ACR : Account Concentration Ratio)를 사용하는데, 지점관여도는 대상기간 중 매수(매도)부분의 상위 다수지점 누적관여율을 나타내는 매수(매도)관여율의 평균값을, 그리고 계좌관여도는 대상기간 중 매수(매도)부분의 상위 다수계좌 누적관여율을 나타내는 매수(매도)관여율의 평균값

을 사용한다.

현재 증권거래소에서 적용하고 있는 기간주가감시 모형은 주가기준 모형과 거래량기준 모형의 경우 시장주가감시 모형과 마찬가지로 평균조정모형(유의수준 5%와 4%)과 2요인 시장모형(유의수준 1%와 2%)을 사용하고 있다. 한편 지점관여도의 경우에는 상위 5개 지점(유의수준 25%), 계좌관여도의 경우에는 상위 20개 계좌를 사용한다.

2.2 증권거래소의 이상매매 적출모형에 대한 평가

현재 증권거래소의 이상매매 적출모형은 간단하고 사용상 편리하다는 장점이 있으나, 다음과 같은 부분에서는 다소의 개선이 필요한 부분도 존재하는 것으로 여겨진다.

2.2.1 통계적 회귀분석 방법론

현재의 이상매매 적출모형은 통계적 회귀분석 모형에 의존함에 따라 적출상황과 불공정 거래행위와의 직접적·체계적 연계성을 파악하는 과정에서는 어려움이 있다. 일반적으로 통계적 방법은 정상적으로 추정되는 수준에서 벗어난 이상거래 종목(outlier)들을 적출하는 데에는 효과적이지만, 적출된 이상거래가 불공정행위가 존재하지 않는 경우 비정상적인 시장조건에 의해서 발생한 결과일 수도 있고 정상적인 시장조건에서의 불공정행위에 의해서 발생한 결과일 수도 있다는 점에서 이들을 서로 구분하는 것이 쉽지 않다는 한계가 있다. 따라서 이상거래 종목들이 모형을 통해 적출된 후에도 명백한 불공정행위가 있었는지의 연계성을 조사·분석하기 위해 상당히 많은 관련 정보들을 추가적으로 검토해야만 필요가 있다.

또한 여러 가지 주식시장 및 기업관련 변수들 가운데 이상매매를 적출하기 위해서 어떤 변수들이 모형에 포함되어야 하며, 어떤 변수를 우선적으로 살펴보아야 하는지 등 변수간의 우선순위에 대한 정보를 과거자료 및 지식을 통해 얻는 것이 용이하지 않다. 이는 통계적 회귀모형이 과거 자료의 검토를 통해서 라기보다는 사전적으로 모형에 사용될 변수들을 선정하고, 설명변수들과 종속변수와의 평면적 관계를 설명하는 모형이기 때문이다.

2.2.2 평균조정모형과 시장모형

현재 거래소의 이상매매 적출모형에서는 주가, 거래량의 예상외 변동을 1차적으로 살펴보고, 혐의종목들에 대해서 시장의 미시구조 자료, 기업의 자본구조, 재무제표자료, 계좌관련 자료 등과 같은 변수들은 종합감리시스템으로부터 추출한 후 분석하는 이원적 구조로 이루어져 있다. 그런데 이와 같은 이원적 구조는 과거의 이상매매 적출과 심리과정에서 얻은 경험과 지식을 현재와 미래의 이상매매를 적출하는 초기 변수로서는 효율적으로 활용하지 못하는 문제를 안고 있다.

또한 현재의 시스템은 불공정거래 행위에 따라 특이하게 나타날 수도 있는 주가수익률과 거래량의 시계열적인 특성을 반영하는 데에는 다소의 어려움이 있다. 물론 현행 시스템에서 기간감시 모형을 통해 이 문제를 보완하고 있으나 변수들의 동태적 측면을 반영하기에는 여전히 한계가 있다고 여겨진다.

2.2.3 이상매매적출 및 심리를 위한 독립적 데이터웨어하우스의 구축

현재 거래소에서는 이상매매 적출모형과 종합감리시스템 등을 통해 특정 종목과 관련된 시장자료들을 얻을 수는 있으나, 기존의 적출상황과 심리

를 통해 얻어진 전문지식을 차후의 적출 작업에 활용하는 것이 어렵고 또한 적출상황이 발생할 때마다 프로젝트 식으로 필요한 데이터와 변수를 새로이 마련하여 분석해야 하는 불편함이 있다. 따라서 이상매매 적출 작업의 효율성을 높이기 위해 독립적인 데이터웨어하우스의 구축을 검토할 필요가 있다고 판단된다.

2.3 NASD의 ADS

2.3.1 ADS의 개요

미국의 NASD(National Association of Securities Dealers)는 1997년 중반부터 '지식발견 및 데이터마이닝 시스템' (Knowledge Discovery and Datamining (KDD) System)을 응용한 ADS를 이상매매의 적출에 활용하고 있다(Kirkland at al., 1999; Senator, 2000). NASD는 NASDAQ을 비롯한 관할 증권시장들을 대상으로 거래, 호가, 주문 자료들로부터 관심대상이 되는 패턴들을 발견하고 불공정매매를 감지할 목적으로 ADS를 사용하는데, 구체적으로, 규제 측면에서 이미 승인된 패턴들로 구성된 지식을 기초로 하여 불공정행위를 야기하는 상황들을 적출하고, 이를 시장분석전문가에게 전달하여 조사하게 한 후 필요에 따라서는 적절한 규제조치를 취한다.

ADS는 주로 다음에 나열한 세 가지 영역에서의 시장조성자(market maker)들의 불공정행위를 적출하기 위해 사용된다.

- 거래보고지연(late-trade reporting)

시장거래자에게 적기에 정확한 정보를 제공하기 위해 모든 거래는 체결 90초 이내에 보고되어야 하는데, 시장조성자가 이것을 의도적으로 여러 번에 걸쳐 위반한 경우

- 시장충실도(market integrity)

시장의 충실도는 시장조성자들 사이의 자유롭고 공정한 경쟁을 전제로 하는데, 불공정한 담합이나 반경쟁적인 행위를 통해 이를 위반하는 경우

- 최선거래체결(best execution)

최선거래체결의 규정이란 투자자는 현행 시장상황 하에서 가장 호의적인 가격조건으로 거래를 체결해야 한다는 것인데, 시장조성자에 의해 이것이 위반된 경우

2.3.2. ADS의 특징

ADS는 한 개의 시스템 내에 적출 및 발견 구성요소를 결합하여 동일한 시장테이터를 공유하면서도 여러 개의 감독영역을 동시에 지원할 수 있다. 또한 ADS는 시각화, 패턴인식 및 데이터마이닝의 다양한 인공지능기능을 활용하여 규제 분석, 패턴감지, 지식발견 등의 활동을 지원한다.

한편 ADS는 '룰 대응'과 '시퀀스 대응'을 주요한 적출기법으로 사용하며, 데이터마이닝 기법을 통해 규제의 관심대상이 되는 새로운 패턴도 발견한다. 또한 ADS는 자료들 사이의 연계성을 찾는데 도움이 될 수 있도록 적절한 기술통계량의 구성, 데이터의 구분, 데이터의 시각화, 대량데이터 환경하에서 시계열적 패턴의 인식 등을 추구하고 있다. ADS 이전에는 시장분석가들이 정보자료를 '테이블 형태'(table format)로 검토함으로써 자료들 간의 연계성을 찾아내기가 힘들었으나, ADS를 통해 이와 같은 문제들이 상당부분 해결될 수 있었다.

ADS는 NASD의 성공사업으로서 NASDAQ 시장감시기능의 효율성을 크게 증가시켰고, 담당부서는 자신의 기능을 안전하고 효과적으로 수행할 수 있게 되었다고 평가되고 있다.

2.3.3 NASD의 ADS가 국내 이상매매 적출모형에 갖는 시사점

앞에서 살펴본 바와 같이 NASD의 ADS가 데이터마이닝 기법을 활용하여 불공정거래의 적출 성과를 높이고 있다는 점을 감안해 볼 때 데이터마이닝 기법의 국내 거래소 이상매매 적출업무에의 도입여부를 검토할 필요가 있다. 아울러 데이터마이닝 기법을 적용한 모형을 도입하는 경우 기존의 적출모형과의 관계를 어떻게 정립해야 할 것인가의 문제도 검토해야 한다. 그러나 이 때 한 가지 염두에 두어야 할 사항은 ADS는 주로 미국 증권시장에서 시장조성자들의 불공정거래를 적출할 목적으로 만들어져 소수행위자의 반복적인 불공정행위를 살펴보기 때문에 데이터마이닝 기법의 활용이 의미가 있을 수 있으나, 국내의 경우에는 불특정 다수투자자들의 특이한 행동을 대상으로 하기 때문에 ADS에서 중요한 역할을 하는 패턴과 시나리오의 형성이 상대적으로 용이하지 않을 수 있다는 점이다.

둘째, ADS는 자체 데이터웨어하우스를 기초로 이상매매 적출에 필요한 주요통계량을 구성하고, 데이터를 그룹별로 구분하거나 시각화하며, 대량 데이터 환경 하에서 시계열 패턴을 인식하고자 하는 등의 지식발견과 데이터마이닝을 응용한 다양한 지원 작업을 수행하고 있는데, 국내에서도 이상매매 적출과 심리에 필요한 데이터웨어하우스를 구축할 필요가 있다고 본다. 특히 지금까지 국내에서 데이터베이스를 구성할 때 전혀 고려대상이 되고 있지 않은 메타데이터, 즉, 적출상황의 감지와 발견 작업에 사용되는 파라미터, 룰패턴과 시퀀스 패턴, 적출 및 발견 작업의 결과물, 다양한 사용자 인터페이스 등을 포함하는 데이터웨어하우스를 구축하고 활용하는 방안을 검토해야 한다.

셋째, ADS는 이상매매의 적출과정에서 시퀀스 매처를 사용하여 자료들의 시계열 관계를 파악하고 있으며 이것이 데이터마이닝 작업의 중요한 부분을 구성하고 있는데, 국내에서도 이와 유사한 형태로 자료들의 시계열 특성을 반영하는 방법을 모색해야 할 것으로 보인다. 즉, 현재 거래소의 적출모형은 자료들의 시계열 특성을 살펴보는데 기간모형에 의존하고 있어서 불공정거래에 따른 관련 변수들의 동태적 특성을 충분히 반영하지 못하므로 이에 대한 보완이 필요할 것으로 생각한다.

3. 이상매매 적출을 위한 마이닝 기법의 적용

데이터마이닝의 관점에서 이상매매 적출의 핵심은 분류(classification)와 예측(prediction)이라고 할 수 있다. 예측과 분류는 어떤 측면에서는 구분하기 곤란할 정도로 유사하다고 할 수 있는데, 일반적으로 이상적인 모형은 현재까지의 주식 관련 각종 자료를 이용하여 이상매매가 발생할 가능성을 비교적 정확하게 예측하여야 하며, 또한 이상매매 종목과 그렇지 않은 종목을 비교적 정확하게 분류할 수 있어야 한다.

따라서 이상매매의 적출에 적합한 데이터마이닝 기법의 후보군은 전통적 통계분석, 신경망과 의사결정나무 정도라고 할 수 있는데, 이와 같은 기법들은 모두 지도학습의 범주에 속한다. 지도학습에서와 같이 과거의 사례와 자료를 이용하여 모형을 훈련시킨다면, 유사한 유형의 불공정매매는 어느 정도 효율적으로 포착할 것으로 예상된다.

그러나 데이터마이닝은 과거의 자료를 이용하여 그 속에 감추어진 패턴을 찾는 것을 목적으로 하는 것이기 때문에 기본적으로 새로운 형태의 이상매매를 찾아내어 이를 분류하는 목적으로 사용되는 데에는 한계가 있다. 물론 연관분석을 통해

각 입력변수들 사이의 연관관계를 어느 정도는 찾아낼 수 있으나, 이는 이상매매 유형의 분류작업에는 크게 도움이 되지 않는다고 볼 수 있다. 자율분류에 속하는 군집분석의 경우도 여러 가지 주식종목을 주어진 수의 군집으로 분류하는 것은 가능하지만, 이상매매 관련 종목들을 행위유형별로 군집화하여 그들의 특성을 살펴보는 작업은 상당히 어려울 것으로 판단된다.

이와 같은 점들을 고려하여 본 연구에서는 지도학습 기법을 우선 적용하기로 하고, 그 가운데 전통적 통계분석 방법으로서 로짓 모형, 그리고 대표적인 지도예측 모형으로서 신경망과 의사결정나무 기법을 적용해 본다. 한편, 이상매매 종목의 경우, 변수들 간의 규칙을 찾아내기 위해 연관분석도 실시하고, 연관분석을 앞의 지도학습 기법들과 연계하여 사용하는 방법도 검토한다. 이는 이상매매로 판정된 종목에 대해 연관분석을 실시하는 경우, 이상매매 종목들에서 특이하게 나타나는 변수간의 관계를 찾아낼 가능성이 충분히 있다고 여겨지기 때문이다.

본 연구에서는 시세조종 종목의 적출에 데이터마이닝 기법을 적용하기 위해 표본을 구축하고 모형설정을 위한 변수를 선정하였다. 이를 위해 우선 최근의 시세조종 종목을 조사하였으며, 비교 및 분석을 위해 같은 기간 동안에 거래된 종목 가운데 정상 종목을 선정하였다.

표본 가운데 시세조종 종목을 선별해내기 위해 시세조종 종목이 갖는 특성을 잘 나타낸다고 생각되는 입력변수를 크게 일일자료, 일중자료 및 기타자료로 구분하여 선정하였다. 변수의 선정에 있어, 우선 시세조종 종목의 특성이 나타나리라고 예상되는 변수를 모두 포함시켜 리스트를 만들었다.

한편, 이상매매 적출을 위한 데이터마이닝 모형 구축에서 자료와 관련된 문제점 가운데 하나는 주식 관련 자료는 패널 데이터이지만 이를 데이터마이닝에 적용하기가 곤란하다는 점이다. 즉, 주식 한 종목이 하나의 레코드(record)를 구성하는데, 주식 한 종목은 시계열 자료로 구성되어 있어 이를 데이터마이닝 모형에 직접 적용하기가 곤란하다는 것이다. 이 문제를 해결하기 위해 여기서는 시계열 자료를 몇 개의 변수로 변환을 하여 사용하였다. 예를 들어, 수익률 시계열 자료의 경우, 1일 수익률, 5일 수익률, 30일 수익률 등을 계산하여 각각 하나의 변수로 이용하였다. 이와 같이 일일자료들의 관측기간을 5일, 30일 등 기간단위로 묶어서 새로운 변수를 만들어 내는 이유는 이들 변수들이 시계열 자료가 갖고 있는 동태적 특성을 어느 정도는 반영해 줄 수 있을 것으로 여겨지기 때문이다. 한편 일중자료는 일일자료와 대응시키기 위해 일중 발생한 여러 가지 거래 및 호가자료를 지표화해서 일일 1회의 관측치로 전환하여 사용한다. 5장에서는 우선 표본을 구성한 후, 표본에 대해 이와 같은 방식으로 선정된 변수들에 대한 사전적 통계분석을 실시한다.

3.1 시세조종 종목과 정상 종목의 선정

본 연구의 표본은 시세조종 종목군과 정상 종목군으로 구분되어 설정된다. 시세조종 종목은 2000년부터 2003년까지의 기간동안 증권거래소에서 적발한 116개 종목으로 구성되며, 정상 종목군은 시세조종 종목군과 비교하기 위해 선정된 종목으로 구성된다. 정상 종목군의 표본은 2000년부터 2003년까지의 기간 동안 거래소에서 거래된 모든 종목 가운데 일부 종목을 추출하여 구성하였다. 정상종목은 총 표본 가운데 위의 기간 동안 시세조종이 없었던 모든 종목과 시세조종 종목이더라도 시세조종 기간이 아닌 기간 중에서 60 거래일의 표본을 선정할 수 있는 종목들 가운데에서 임의로

선정하여 총 844 종목을 얻었다.[1] 그러나 임의로 선정된 기간 중에는 특정 주식의 거래가 이루어지지 않았던 종목들이 있어 통계적 분석의 문제[2]가 발생하므로 6장의 실증분석에는 이들을 제외한 670개의 종목을 사용하였다. 한편, 일중자료(intra data)를 사용하는 경우에는 670개의 종목 가운데 다시 200개 종목을 임의표본으로 선정하여 정상종목군을 선정하였다. 일중자료를 사용함에 있어 정상 종목군의 표본수가 적게 선택된 것은 일중자료에서는 시세조종의 패턴을 잘 살펴보기 위해 시세조종 종목들이 시세조종 기간이 아닌 다른 기간에서 보여 준 거래량이나 시가총액 등과 유사한 종목으로만 정상 종목으로 구성했기 때문이다. 한 가지 예로서 본 연구에서는 시가총액 하위 400위 안에 포함된 비교적 소형종목만으로 정상 종목군을 선정하였다. 한편 본 연구에서 일중자료를 활용한 분석을 제외한 모든 분석은 670개의 종목으로 구성된 정상 종목군을 중심으로 수행되었다.

이렇게 종목의 선정을 마친 후 해당 종목의 관측치들은 각 종목의 마지막 거래일을 기초로 산출되었다. 예를 들어, 시세조종 종목의 경우 시세조종기간이 2001년 2월 3일부터 2001년 5월 6일까지라면 마지막 날인 2001년 5월 6일을 기준으로 분석대상이 되는 변수의 값이 계산되었다. 즉, 분석변수가 30일 누적수익률이라면 2001년 5월 6일을 포함하여 이전 30일 동안 누적한 수익률로 계산된다. 한편 정상종목군의 경우에는 선정된 기간의 마지막 날짜를 기준으로 변수들의 값이 계산되었다.

1) 이에 따라 같은 종목이라 하더라도 상이한 거래 기간의 종목으로 두 번 이상 표본에 포함이 될 수도 있다. 표본의 되는 종목의 자료를 60거래일씩 추출한 것은 시세조종 종목의 시세조종 기간이 일정하지는 않으나 대개 60 거래일 정도에 시세조종 기간이 충분히 포함되기 때문이다.

2) 추후 살펴볼 초과수익률과 초과 거래회전율을 얻기 위해서는 충분한 계수 추정기간이 필요하다. 그러나 본 연구에서 우선주의 경우에는 주어진 기간동안 거래가 전혀 없던 경우도 빈번하여 적절한 계수 추정치를 얻지 못하는 경우가 많이 발생하였고, 따라서 이 경우에는 결손치(missing value)로 처리하였다.

3.2 데이터마이닝 변수의 선정

데이터마이닝의 경우에는 변수들의 선택이 다른 통계적 분석들과 비교하여 상대적으로 자유롭다. 물론 관측치의 수에 비해 변수가 과도하게 많이 설정되는 경우 앞장에서 언급한 차원의 저주라는 문제가 발생할 수 있으나, 일단은 데이터마이닝 작업에 사용할 변수를 선정함에 있어서 중요하다고 생각하는 변수들은 모두 포함하기로 한다. 구체적으로 모형에 포함될 입력변수들은 크게 다섯 가지로 나누어지는데, 첫째는 종목의 수익률과 거래량 자료, 둘째는 자료의 시계열 특성을 파악하기 위한 자기상관 관련 자료, 셋째는 계좌관여도와 지점관여도, 넷째는 기업의 재무성과 자료, 그리고 마지막으로 종목의 일중자료들이다. 이제 이들 변수들에 대해서 보다 자세히 살펴본다.

3.2.1 초과수익률

초과수익률은 종목의 실제수익률에서 기대수익률을 차감한 것으로 정의되며, 본 연구에서는 기대수익률을 1년 수익률 평균과 시장모형을 이용하여 다음과 같이 계산한다.

- 1년 평균수익률을 사용하여 기대수익률을 계산하는 경우

$$ER_{i,t} = R_{i,t} - \overline{R}_i$$

- 시장모형을 사용하여 기대수익률을 계산하는 경우

$$ER_{i,t} = R_{i,t} - \alpha_i - \beta_i R_{m,t}$$

시장모형에서의 모수 추정과 평균수익률의 계산은 현재에서 1년 3개월 전부터 3개월 전까지의 1년 자료를 사용하여 추정하였다. 이와 같이 계산된 초과수익률을 1일, 1~8주, 10주, 3개월로 누적하여 다음과 같이 누적 초과수익률 변수를 정의한다.

$$ER_i(d,\ t)\ =\ \sum_{k\ =\ -d+t+1}^{t} ER_{i,\ k}$$

여기에서 d = 1, 5, 10, 15, 20, 25, 30, 35, 40, 50, 60으로 설정하여 각각의 경우에 누적 초과수익률을 계산한다. 그런데 해당 일수는 거래일을 나타내기 때문에 d = 10은 달력일(calendar date)로는 2주를, d = 40은 2개월을 의미한다.

누적 초과수익률을 계산하여 비교해본 결과, 어느 모형을 사용하여 기대수익률을 계산하는지와 상관없이 15일 이상부터는 시세조종 종목의 누적 초과수익률이 정상 종목의 누적 초과수익률보다 유의적으로 높음을 알 수 있었으며, 이러한 결과는 시세조종 여부를 판단하는 중요한 변수로서 수익률 자료를 사용하는 것이 중요하다는 것을 지지하는 증거라고 할 수 있다.

3.2.2 초과 거래회전율

초과 거래회전율은 실제 거래회전율에서 기대 거래회전율을 차감한 것으로 정의되며, 본 연구에서는 기대 거래회전율을 1년 평균 거래회전율과 시장모형을 이용하여 다음과 같이 계산한다.

- 거래회전율의 정의

  $LTO_{i,\ t} = \log(\ V_{i,\ t}) - \log(\text{발행주식수}_{i,\ t})$

- 1년 평균 거래회전율을 사용하여 기대 거래회전율을 계산하는 경우

  $ELTO_{i,\ t}\ = LTO_{i,\ t}\ -\ \overline{LTO_i}$

- 시장모형을 사용하여 기대 거래회전율을 계산하는 경우

  $ELTO_{i,\ t}\ = LTO_{i,\ t}\ -\ \alpha_i - \beta_i LTO_{m,\ t}$

- 시장 거래회전율

  $LTO_{m,\ t} = \sum w_i LTO_i$

  (ω_i는 발행주식수 기준의 가중치)

시장모형에서의 모수 추정과 평균 거래회전율의 계산은 현재에서 1년 3개월 전부터 3개월 전까지의 1년 자료를 사용하여 추정하였다. 이와 같이 계산된 초과 거래회전율을 1일, 1~8주, 10주, 3개월로 누적하여 다음과 같이 누적 초과 거래회전율 변수를 정의한다.

$$ELTO_i(d,\ t)\ =\ \sum_{k=-d+t+1}^{t} ELTO_{i,\ k}\, z$$

여기에서 d = 1, 5, 10, 15, 20, 25, 30, 35, 40, 50, 60으로 설정하여 각각 누적 초과거래회전율을 계산한다.

누적 초과 거래회전률을 계산하여 비교해본 결과, 기대 거래회전율을 어느 모형으로 사용하여 추정하는가 하는 것과 추정기간에 상관없이 시세조종 종목군의 누적 초과 거래회전율이 정상 종목의 누적 초과 거래회전율보다 유의적으로 크다는 사실을 발견할 수 있다. 그런데 초과수익률과는 달리 그 규모의 차이가 장기 누적 초과거래회전율보다는 단기 누적 초과거래회전율에서 크게 나타남을 알 수 있었다. 이러한 현상은 시세조종 마지막일 직전에 거래량이 크게 증가했기 때문에 발생한 것으로 추측된다.

3.2.3 자기상관계수

종목들의 가격변화 혹은 수익률의 시계열 특성

을 반영하기 위해 다음과 같이 정의되는 자기상관계수(autocorrelation)를 변수로 도입한다.

$$\rho_i(d,\ t) = \sum_{k=-d+1+t}^{t} R_k R_{k-1} \Big/ \sum_{k=-d+1+t}^{t} R_k^2$$

본 연구에서는 $d = 5$로 고정하고 수익률을 일별 수익률과 주간수익률로 나누어 일별수익률의 5일간 자기상관계수와 주간수익률의 5주간 자기상관계수를 계산하였다. 이러한 단기 자기상관계수는 시계열에 대한 시스템적 해석을 위해서라기보다 단기적으로 나타나는 시계열의 특성을 파악하기 위해서이다.

한편, 시세조종 마지막일을 기준으로 3개월 전까지의 수익률과 거래회전율 자료를 이용하여 수익률과 거래회전율에 대한 AR(1), AR(2), AR(3) 계수를 각각 산출하여 수익률과 거래회전율의 시스템적인 시계열 특성을 파악한다.

자기상관계수를 계산하여 비교해본 결과, 시세조종 종목군의 경우 정상 종목군보다 대체로 자기상관성이 높음을 알 수 있다. 그리고 시세조종 종목군의 경우 주로 양(+)의 자기상관관계를 갖고 있는 특징을 보이고 있다. 예를 들어, 5일간의 자기상관계수를 보면 시세조종 종목군의 경우 0.18의 높은 양의 자기상관관계를 갖고 있으나, 정상 종목군의 경우는 0에 가까운 자기상관관계를 가지는 것으로 나왔다. 일반적으로 주식시장의 효율성을 측정할 때 자기상관관계가 0인 것을 가정하고 있는 것으로 볼 때 이러한 자기상관적 특징은 시세조종 종목을 선별하는 중요한 지표로서 활용할 수 있을 것으로 생각된다.

3.2.4 재무비율

시세조종 종목들의 재무적 특성을 살펴보기 위해 다음과 같은 재무지표들은 고려한다.

- 직전년도의 자본금 : 보통주와 우선주의 자본금 규모
- 직전년도의 ROE(당기순이익/자본총계), ROA(당기순이익/자산총계)
- 직전년도의 부채비율(부채총계/자본총계)

기업의 자본금 규모는 자본금의 절대금액에 로그를 취한 후 이를 정규화한 값을 사용하였다.

그 결과 자본금과 ROA의 경우 시세조종 종목군과 정상 종목군이 유의적인 차이를 보이며, 특히 자본금 규모의 경우에 시세조종 기업들의 특징이 잘 드러나 있는 것으로 판단된다.

3.2.5 관여도

거래일 동안 투자자 거래행태의 특성을 반영하기 위한 변수로 지점관여도와 계좌관여도를 사용한다. 지점관여도(branch concentration ratio)는 대상기간 중 매수부분의 상위 다수지점 누적관여율과 매도부분의 상위 다수지점 누적관여율을 평균한 값이며, 계좌관여도(account concentration ratio)는 대상기간 중 매수부분의 상위 다수계좌 누적관여율과 매도부분의 상위 다수계좌 누적관여율을 평균한 값이다. 관여도 자료는 2001년 5월 9일부터 누적되어 있기 때문에 이전 기간은 결손치로 처리하였다. 관여도는 이동평균의 개념으로 계산되었으므로 마지막 날의 지점 및 계좌관여도와 5일전, 10일전 지점 및 계좌관여도를 사용하여 시계열 특성을 반영하고자 하였다.

그 결과, 시세조종 종목군의 관여도는 정상 종목군과 비교할 때 유의적인 차이를 보이지는 않았다. 이는 사용한 관여도 값이 이동평균에 의해 구해진 값이기 때문에 기간 누적에 의하여 발생한 것으로 추정되지만 이에 대한 확정적인 증거는 제시되고 있지 않다.

3.2.6 기타

본 연구에서 사용된 나머지 일일자료 변수들은 다음과 같다.

- 누적수익률

5일, 10일, 20일, 30일, 40일, 50일, 60일 간의 누적수익률

- 거래회전율

10일, 20일, 30일, 40일, 50일, 60일 간의 평균 거래회전율

- 일중 스프레드

(최고가-최저가)/최저가의 5일, 10일, 20일, 30일, 40일 평균값

- 최고가 대비 현재가격

시세조종 시작 전 1년 동안의 최고가와 시세조종 마지막 날의 주가와의 비율

- 수익률 변동성

수익률의 20일, 40일 표준편차

- 거래회전율 변동성

거래회전율의 20일, 40일 표준편차

3.2.7 일중자료(intra-day data)

일중자료는 크게 취소정정 건수, 대량매매, 매도/매수 비율, 매매구분, 불이익주문, 스프레드 등의 항목으로 나누어 정리하였다. 각 항목들에 대한 설명과 아울러 관련 된 변수들의 이름을 나열하면 다음과 같다.

- 주문건수(주문량) 대비 취소정정 건수(주문량)

매도/매수주문별 주문건수(주문량) 대비 취소와 정정 건수(주문량)의 10일 평균값과 동 평균값 대비 마지막 거래일 값의 비율

- 주문건수 대비 체결건수

주문건수 대비 체결 건수의 10일 평균값과 동 평균값 대비 마지막 거래일 값의 비율

- 매도주문 대비 매수주문

매도주문 건수(주문량) 대비 매수주문건수(주문량) 비율의 10일 평균값과 동 평균값 대비 마지막 거래일 값의 비율

- 불이익 주문

매도(매수)의 경우 불이익 주문은 매수(매도)우선호가보다 낮게(높게) 매도(매수)주문을 낸 경우를 의미함. 매도(매수)의 경우 전체 매도(매수)주문 건수(주문량) 대비 불이익 매도(매수)주문 건수(주문량)의 비율의 10일 평균값과 동 평균값 대비 마지막 거래일 값의 비율

- 대량매매

전체 주문건수(주문량) 대비 대량매매 건수(주문량), 전체 체결건수(체결량) 대비 대량매매 체결건수(체결량)의 10일 평균과 동 평균값 대비 마지막 거래일 값의 비율

- 매매 구분

시가, 접속, 종가, 시간외 주문량을 전체 주문량으로 나눈 비율의 10일 평균값과 동 평균값 대비 마지막 거래일 값의 비율

- 스프레드

(최우선 매도호가-최우선 매수호가)/(최우선 매도호가+최우선 매수호가)/2를 계산하여 특정일의 평균값을 측정한 후 그 평균의 10일 평균값과 동 평균값 대비 마지막 거래일 값의 비율

4. 데이터마이닝을 이용한 이상매매 적출 분석

4.1 모형의 선택과 적용

4.1.1 모형의 선정

본 절에서는 앞에서 소개된 자료와 변수를 바탕으로 여러 가지 데이터마이닝 기법을 이상매매 적출에 적용하도록 한다. 이상매매를 적출해내기 위해서는 우선 이상매매의 대상이 되는 주식의 특성을 파악하여야 하며 이상매매의 경우 나타나는 시계열적 패턴을 찾아내어야 한다. 만일 이상매매의 대상이 되는 주식의 특성과 이들의 시계열적 패턴이 확인된다면, 유사한 특성과 패턴을 보이는 종목을 이상매매 징후가 있는 것으로 판단할 수 있을 것이다. 이 같은 과정에 적합한 방식은 지도학습 방법이다.

지도학습의 문제점은 프로그램을 훈련시키기 위해 방대한 양의 훈련용 데이터가 필요하다는 것이다. 본 연구에서 사용할 수 있는 자료는 이상매매 종목이 116개로 일반적인 데이터마이닝에서 사용되는 관측치에 비해 상당히 작은 편이다. 이 경우에는 소위 차원의 저주(curse of dimensionality) 문제가 발생할 수 있으므로 모형을 주의 깊게 구성할 필요가 있다(Roiger, 2003).

본 연구에서는 지도학습 기법 가운데 로짓(logit), 인공신경망(neural network) 및 의사결정나무(decision tree) 모형 등을 적용하기로 한다. 인공신경망 모형은 패턴 인식에 탁월한 능력을 보이고 있으나 모형을 훈련시키는 데는 방대한 양의 데이터가 필요하다. 로짓 모형에서는 적정한 관측치 수가 얼마가 되어야 하는가에 대한 이론적인 근거는 없으나 충분치 않은 수의 자료를 이용할 경우 추정치의 불편성과 효율성에 문제가 발생하여 그 결과를 신뢰할 수 없다. 현재 확보되어 있는 표본의 수를 고려할 때, 로짓 모형과 인공신경망 모형은 이상매매를 선별하는데 효율성이 높지 않을 가능성이 있는 것으로 보인다. 이에 비해 의사결정나무는 차원의 저주 문제로부터 어느 정도 자유롭다. 이는 의사결정나무에는 자체적으로 차원을 축소시키는 메커니즘이 내재되어 있기 때문이다. 따라서 로짓이나 인공신경망에 비해 비교적 적은 수의 자료를 이용하여 모형을 훈련시키는 것이 가능하다. 이제까지 논의한 바를 종합하면, 현 상태에서 이상매매 적출에 가장 효율적인 지도학습 기법은 의사결정나무일 것으로 추정된다. 현재 미국 NASD의 ADS도 의사결정나무를 사용하고 있는데, 이 역시 의사결정나무가 갖는 장점 때문일 것으로 보인다.

4.1.2 변수의 선택

데이터마이닝을 통해 이상매매를 선별해내기 위해서는 먼저 이상매매 종목의 특성을 나타낼 가능성이 있는 변수의 후보를 선정하여야 한다. 데이터마이닝이 직접 이상매매 종목의 특성을 나타내는 변수를 제공해 줄 수는 없으나, 데이터마이닝을 통해 이상매매를 선별해내기 위한 변수를 결정할 수는 있다. 앞에서 제시된 바와 같이 본 연구에서는 데이터마이닝에 이용될 수 있다고 판단되는 변수를 크게 일별 변수와 일중 변수로 나누어 모두 130개를 선택하였다.

변수가 이 같이 많아진 것은 주가관련 자료가 패널 데이터(panel data)라는데 기인한다. 특정 종목의 경우, 특정일에 그 종목은 하나의 관측치가 되며, 그 날에는 상장주식 수만큼의 관측치가 존재한다. 그런데 하나의 관측치에는, 예를 들어 주가라는 변수가 시계열로 존재하며, 호가와 같은 일중 자료도 시계열로 존재한다. 이러한 패널 데이터를 데이터마이닝에 직접 적용하는 것은 해결하기 매우 곤란한 과제이다. 따라서 본 연구에서는 시계열

자료를 하나의 변수로 요약하여 사용하기로 한다. 예를 들어, 수익률이라는 변수는 하나의 시계열인데 이를 전일 수익률, 전5일 수익률, 전10일 수익률, 전20일 수익률 등등으로 전환하여 각각을 변수로 이용하는 것이다.

그러나 이 같은 방식은 관측치에 비해 변수가 과도하게 많아지는 문제가 있다. 따라서 변수의 수를 줄일 필요가 있다. 변수의 수를 축소하기 위한 방편으로 의사결정나무를 이용하기로 한다. 즉, 우선 모든 변수를 포함시켜 의사결정나무 모형을 실행한 후 여기서 남는 변수와 t-검정 결과 유의도가 높은 변수를 선정하여 다시 모형을 실행하여 그 결과를 분석하는 것이다.

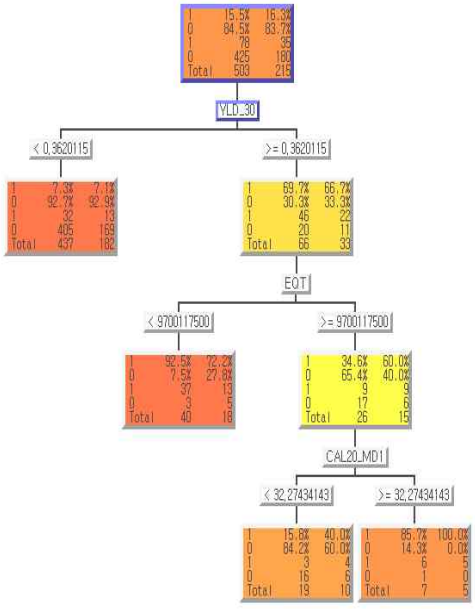
4.2 의사결정나무 결과 분석

본 절에서는 의사결정나무를 실행하여 얻은 결과를 분석하기로 한다. 앞서 논의한 바와 같이 현재 이용가능한 관측치의 수는 많지 않은 반면 변수의 수는 많은 편이다. 또한, 많은 관측치에서 일부 변수가 결손값인 경우가 있었다. 따라서 모형의 성과를 높이기 위해 변수의 수와 결손값 등을 고려하여 모형을 설정하고 추정하기로 한다.

의사결정나무에서 분기의 기준으로 사용되는 통계량은 Chi-square test, Gini reduction, entropy reduction 등이 있으며, 이외에도 분기의 효율성을 제고하기 위해 유전자 알고리듬을 이용하는 등 새로운 방법이 시도되고 있다. 의사결정나무에서 분기를 위한 변수는 분류하는 경우 그 분포가 모집단의 분포와 될 수 있으면 큰 차이를 나타내는 것이 선택되며 이 때 적용되는 것이 앞에서 언급한 통계량이다. 본 절에서는 이 가운데 널리 쓰이는 Chi-square test를 통해 분기 여부를 결정하기로 한다. 한편, 의사결정나무를 모형화하는데 모형의 깊이(분기의 단계 수)와 마디(node)에서의 분기의 수를 사전에 결정할 필요가 있다. 여기서는 분기의 수는 기본적으로 2를 사용하고 경우에 따라 확대하기로 한다. 모형의 깊이는 대부분 6단계 이하인 것으로 나타나고 있어 10단계 이상 내려가지는 않는 것을 기본으로 한다.

4.2.1 일중자료를 제외한 전체 데이터

우선, 모형의 전반적인 성과를 보기 위해 일일 자료 전체를 대상으로 모형을 실행해보기로 한다. 총 관측치는 718개이며 이 가운데 우선주는 129개이다. 전체 관측치 가운데 이상매매 종목은 113개이며 나머지 605 종목이 정상 종목이다. 우선주는 각각 33종목과 96종목이 포함되어 있다. 한편, 모형 실행에 이용된 변수의 수는 모두 86개이다. 모형의 적합성을 판단하기 위해 718개의 관측치를 크게 훈련용 데이터와 검증용 데이터로 나누어 분석을 실행한다. 검증용 데이터는 전체 자료 가운데 30%를 프로그램이 임의로 선택하게 하였다.



[그림 1] 의사결정나무 실행 결과

[그림 1]에는 SAS의 Enterprise Miner를 이용한 의사결정나무 실행 결과가 나타나 있다(SAS Intstitute Inc., 2002). 이에 의하면, 이상매매를 선별하는데 주요 변수는 YLD_30(30일간 수익률), EQT(자기자본), CAL20_MD3(모형 3에 의한 20일간 초과거래량) 등이다. 그림에서 각 마디의 상자는 그 안에 포함되어 있는 자료의 구성을 의미하고 있다. 첫 번째 열에서 1은 이상매매 종목을 의미하며 0은 정상종목을 의미한다. 두 번째 행은 훈련용 데이터, 그리고 세 번째 행은 검증용 데이터에 각각 포함된 이상매매 종목과 정상 종목의 비중 및 수를 나타낸다. [그림 1]에 의하면 훈련용 데이터에서는 90%의 성공률로 이상매매를 분류해냈으나 검증용 데이터에서는 성공률이 60%대로 떨어지는 것으로 나타났다.

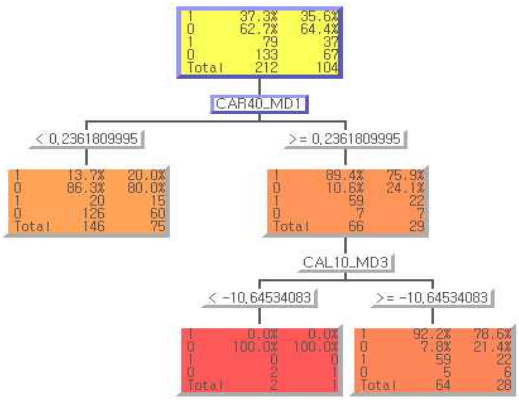
<표 1>에는 위에서 수행된 의사결정나무의 오분류행렬이 제시되어 있다. 표에 의하면 전체 215개의 검증용 자료 가운데 24개가 이상매매로 분류되었다. 이 가운데 정상 종목을 이상매매로 분류한 것이 5종목이다. 한편 이상매매종목 16개를 정상종목으로 분류하고 있어 적중률은 50% 수준인 것으로 나타났다. 이 같은 결과는 현재의 의사결정나무의 성과가 그다지 좋지 못하다는 것을 보여주고 있다. 이는 우선 변수의 수가 너무 많고 다음으로 결손값이 많기 때문에 나타난 결과로 보인다. 특히 의사결정나무에서는 결손값도 의사결정의 요소로 이용하기 때문에 결손값이 결과를 왜곡시킬 가능성이 있다.

<표 1> 의사결정나무의 오분류행렬

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	175 81.40 97.22 91.62	5 2.33 2.78 20.83	180 83.72
1	16 7.44 45.71 8.38	19 8.84 54.29 79.17	35 16.28
총 합	191 88.84	24 11.16	215 100.00

4.2.2 일중자료를 포함한 전체 데이터

일중자료를 포함한 전체 데이터세트는 모두 316개의 관측치로 구성되어 있다. 이 같이 관측치가 줄어든 것은 일중자료 관련 변수를 일부 관측치에 대해서만 구축하였기 때문이다. 316개의 관측치 가운데 정상 종목은 200개이며 이상매매 종목은 116개이다. 이 가운데 우선주는 정상종목에 34개, 그리고 이상매매 종목에 34개가 각각 포함되어 있다. 총 변수는 모두 130개로서 이 가운데 86개가 일일자료 및 기타자료를 이용하여 구축되었고 46개 변수가 일중자료를 이용하여 구축되었다. 그 결과는 [그림 2]에서 보듯이 매우 간단하다. 이상매매 적출에 이용된 변수는 40일 초과수익률(CAR40_ MD1)과 10일 초과거래량(CAL10_md3)뿐이다.



[그림 2] 의사결정나무 : 일중자료를 포함한 경우

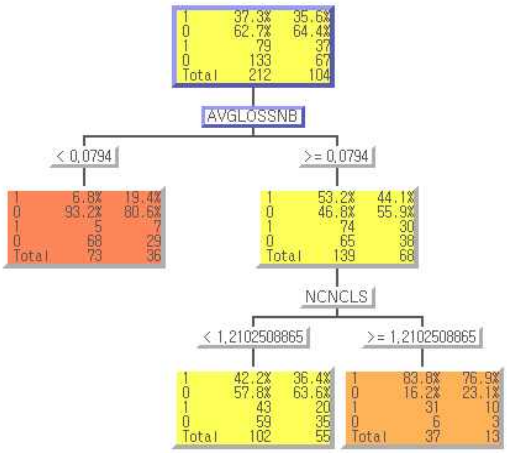
그러나, 오분류행렬을 살펴보면 모형의 성과는 오히려 앞에서 제시된 여러 변수를 포함하고 있는 모형보다 우수한 것으로 나타났다. 정상종목 6개를 이상매매 종목으로 분류하기는 하였으나 37개의 이상매매 종목 가운데 22개를 제대로 선별해냈다. 이와 같은 결과는 이상매매 종목의 특성은 역시 수익률과 거래량에서 나타난다는 것을 의미한다.

<표 2> 의사결정나무의 오분류행렬 : 일중자료 포함

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	61 58.65 91.04 80.26	6 5.77 8.96 21.43	67 64.42
1	15 14.42 40.54 19.74	22 21.15 59.46 78.57	37 35.58
총 합	76 73.08	28 26.92	104 100.00

4.2.3 일중자료 관련 변수만 적용된 경우

이 번에는 변수 가운데 일중자료와 관련이 없는 86개의 변수를 제외하고 모형을 설정하기로 한다. 관측치는 앞에서의 경우와 동일한 반면 변수는 일중자료와 관련된 것 44개만 포함되었다. 그 결과는 [그림 3]에 도시되어 있다. 그림에 의하면 일중자료 관련 변수 가운데 유의하게 이용된 것은 ABGLOSSNB (매수불이익주문건수/매수주문건수 10일 평균)와 NCNCLS(매도취소정정건수/매도주문건수 10일 평균 대비 최종일 값)이다. 그러나 이러한 변수만을 사용하였을 경우, 성과는 좋지 않은 것으로 나타났다. 한편, 변수 선정 변화와 성과 향상 여부를 보기 위해 본 자료의 경우에는 각 마디에서의 분기가 3번 이상이 허용되도록 하여 다시 모형을 설정하였다. 이러한 설정에서는 취소정정량/전체주문량 10일 평균, 매수주문수량/매도주문수량 10일 평균, 매수불이익주문건수/매수주문건수 10일 평균, 시간외 주문량/전체주문량 10일 평균 등이 포함되었으나 오분류 행렬상에서 유의한 향상은 보이지 않았다.



[그림 3] 의사결정나무 : 일중자료 변수만 이용된 경우

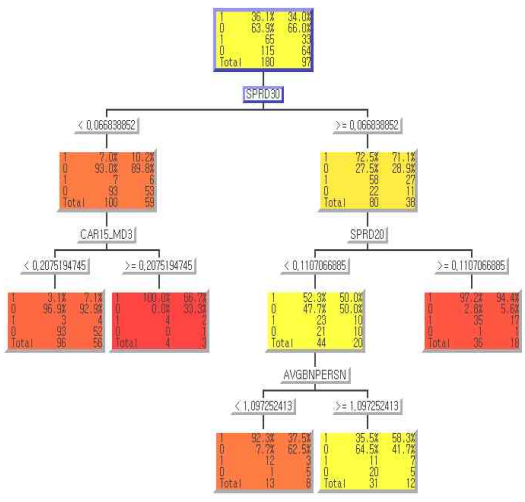
<표 3> 의사결정나무의 오분류행렬 : 일중자료 변수만 이용된 경우

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	64 61.54 95.52 70.33	3 2.88 4.48 23.08	67 64.42
1	27 25.96 72.97 29.67	10 9.62 27.03 76.92	37 35.58
총 합	91 87.50	13 12.50	104 100.00

4.2.4 모든 변수를 적용한 경우 : 주요 결손치 포함 관측치 제외

여기서는 모든 변수를 포함시켰을 경우 의사결정나무가 어떠한 성과를 보이는가 분석한다. 특히, 앞에서 확인된 바와 같은 결손치로 인한 왜곡을 해결하기 위해 다수의 결손치를 포함하고 있는 관측치는 제외하기로 한다. 이 경우, 총 관측치는 모두 277종목이며 이 가운데 정상은 179종목(우선주 19종목), 그리고 이상매매는 98종목(우선주 29종목)이다. 총변수는 일일자료 및 기타 자료에서 구축된 86개와 일중자료에서 구축된 44개를 포함하여 모두 130개이다.

모형의 수행한 결과는 [그림 4]에 나타나 있다. 이상매매 선별에 사용된 변수는 SPRD30((최고가-최저가)/최저가의 30일 평균), CAR15_MD3(시장모형에 의한 15일), CAR SPRD20((최고가-최저가)/최저가의 20일 평균) 및 AVGBNPERSN(매수주문건수/매도주문건수 10일 평균) 등으로 일별 및 일중 수익률 관련 변수가 주로 선택되었다.



[그림 4] 모든 변수가 이용된 경우 : 다수의 결손치가 있는 관측치 제외

한편, 오분류행렬에 의하면 33개의 이상매매 종목 가운데 67% 가량인 22개 종목을 제대로 분류해낸 것으로 나타났다. 이와 같은 성과는 이제까지의 모형 가운데 가장 우수한 것이다. <표 5>에서는 의사결정나무에서 적용된 네 종류의 입력 데이터를 사용한 모형의 적중률(이상매매를 이상매매로 분류한 비율)을 훈련용 데이터와 검증용 데이터에 대하여 각각 보여주고 있다.

<표 4> 의사결정나무의 오분류행렬

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	57 58.16 89.06 82.61	7 7.14 10.94 24.14	64 65.31
1	11 11.22 33.33 15.94	22 22.45 66.67 75.86	33 33.67
총 합	69 70.41	29 29.59	98 100.00

<표 5> 네 종류의 입력 데이터를 사용한 모형의 적중률

데 이 터	적 중 률	
	훈련용	검증용
일중자료를 제외한 전체 변수 적용	59%	54%
일중자료를 포함한 전체 변수 적용	75%	59%
일중자료 관련 변수만 적용	39%	27%
모든 변수를 적용	78%	67%

4.3 로짓 모형 분석

로짓 모형은 로지스틱 함수를 이용한 회귀분석 모형으로 종속변수가 확률로 나타난다. 여기서는 이상매매 종목을 1, 그리고 정상 종목을 0으로 하

여 모형을 추정한다. 모형의 계수가 모두 추정된 후 새로운 관측치의 변수값을 대입하면 해당 관측치의 종속변수 추정치가 0에서 1사이 값으로 나오게 된다. 만일 이 값이 1에 가깝다면 이 관측치를 이상매매 종목으로 분류하고 0에 가깝다면 정상종목으로 분류한다. 그러나 이론적으로 확률값이 얼마 이상이 되어야 이상매매 종목으로 분류한다는 기준은 존재하지 않는다.

4.3.1 일중자료를 제외한 변수

우선 일중자료 관련 변수를 제외한 나머지 86개 변수의 계수를 718종목의 관측치를 이용하여 모형을 추정하였다. 자료의 구성은 2절의 가-1)과 동일하다. 그런데, 로짓 분석에서는 다수의 변수 가운데 설명력이 높은 변수를 선별해내는 단계식 회귀분석(stepwise regression)을 실행하는 것이 가능하다. 따라서 여기서는 단계식 회귀를 통해 우선 변수를 선별하여 모형을 추정하기로 한다. 단계식 분석의 결과 채택된 변수는 sprd10((최고가-최저가)/최저가의 10일 평균)과 bcrbf10 (전 10일 지점 관여도)이다. 로짓 모형에 의한 결과는 <표 6>에 제시되어 있는데 거의 이상매매를 선별해내지 못하고 있다.

한편, 동일한 자료 가운데 모형에서의 계수의 수를 줄이기 위해 일부 변수를 제한 모형을 추정하여 오분류행렬을 작성하였는데 그 결과도 앞의 표와 크게 다르지 않았다. 이러한 결과는 결손치가 포함된 관측치를 제거한 표본에서도 마찬가지로 나타났다.

4.3.2 일중자료 관련 변수만 포함된 경우

일중자료 관련 변수의 영향을 알아보기 위해 여기서는 앞에서 소개된 130개의 변수 가운데 44개의 일중자료 관련 변수만 이용하여 로짓 모형을 추정하였다. 그 결과 avgintQ, avglossNS, avgNCncl, bgq, NCnclS 등이 단계식 회귀분석에 의해 선택되었다. 한편, 모형 추정에 의한 오분류행렬은 <표 7>에 제시되어 있는데 효율성이 제한적이기는 하지만 그 결과는 앞의 분석 결과보다 크게 향상된 것으로 나타났다. 이는 적어도 로짓 모형에서는 일중

<표 6> 로짓 모형의 오분류행렬: 일중자료 관련 변수 제외

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	178 82.79 98.89 85.58	2 0.93 1.11 28.57	180 83.72
1	30 13.95 85.71 14.42	5 2.33 14.29 71.43	35 16.28
총 합	208 96.74	7 3.26	215 100.00

<표 7> 로짓 모형의 오분류행렬 : 일중자료 관련 변수만 포함

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	50 52.63 83.33 73.53	10 10.53 16.67 37.04	60 63.16
1	18 18.95 51.43 26.47	17 17.89 48.57 62.96	35 36.84
총 합	68 71.58	27 28.42	95 100.00

자료 관련 변수의 중요도가 클 가능성이 있다는 것을 시사한다.

4.3.3 모든 변수가 포함된 경우

130개의 변수 모두를 포함시켜 로짓모형을 추정한 결과가 <표 8>에 제시되어 있는데, 31개 이상매매 종목 가운데 22개를 제대로 선별하여 70% 이상의 적중률을 보이고 있다. 이와 같은 결과는 결손치가 있는 관측치를 제외한 경우도 유사하였는데, 정상 종목을 이상매매 종목으로 잘못 분류하는 경우가 2건으로 줄어들었다.

<표 8> 로짓 모형의 오분류행렬 : 모든 변수 포함

빈도 백분율 행 백분율 칼럼 백분율	0	1	총 합
0	57 61.29 93.44 85.07	4 4.30 6.56 15.38	6 65.59
1	9 9.68 29.03 13.43	22 23.66 70.97 84.62	31 33.33
총 합	67 72.04	26 27.96	93 100.00

한편, 이 모형에서 선택된 독립변수는 ar_d5, Cncls 및 sprd20이었다. 이는 매도취소정정, 장중 최고가와 최저가 차이 그리고 수익률의 자기상관 등이 이상매매를 선별하는데 주요한 변수라는 것을 의미한다. <표 9>에서는 로짓 모형에서 적용된 세 종류의 입력 데이터를 사용한 모형의 적중률(이상매매를 이상매매로 분류한 비율)을 훈련용 데이터와 검증용 데이터에 대하여 각각 보여주고 있다.

<표 9> 세 종류의 입력 데이터를 사용한 모형의 적중률

데 이 터	적 중 률	
	훈련용	검증용
일중자료를 제외한 전체 변수 적용	25%	14%
일중자료 관련 변수만 적용	60%	49%
모든 변수를 적용	78%	71%

4.4 기타 모형의 적용

4.4.1 인공신경망 모형

인공신경망의 네트워크 아키텍처로는 multilayer percetron을 선택하였으며, 현재 이용가능한 관측치의 수가 제한적이고 투입요소의 수가 많다는 점을 고려하여 은닉층(hidden layer)은 하나로 제한하였다. 은닉층이 많아진다면 추정해야할 가중치의 수가 기하급수적으로 많아지므로 하나 이상의 은닉층을 적용하는 것은 실질적으로 불가능하다고 할 수 있다. 인공신경망을 이용한 분석은 의사결정나무에서와 같이 데이터 세트와 변수를 다양하게 변화시켜 실시하였다.

그 결과 수익률관련 변수만 이용한 경우가 가장 성과가 좋았으며, 이상매매 종목 가운데 제대로 분류된 것이 60%를 상회할 정도의 성공률을 보이고 있다. 이는 이상매매 종목이 역시 수익률에서 정형화된 패턴을 보일 가능성이 크다는 것을 의미한다. 그러나 물론, 동일한 모형이 다른 데이터 세트에서도 성과가 좋으리라는 보장은 없다. 이는 인공신경망을 이용한 주가 예측에서도 나타나는 문제로 일정 기간 동안 인공신경망 모형이 주가를 탁월하게 예측하였지만 곧 예측력이 현저히 감소되는 경우가 있다. 이는 주가 움직임의 패턴이 늘 변화하기 때문인데, 이상매매 종목의 경우도 수익률 움직임 패턴이 시장의 변화에 따라 변화할 가능성이 있다.

따라서 이 모형의 성과를 판단하기 위해서는 보다 광범위한 자료를 이용한 검증이 필요하다.

4.4.2 연관규칙

연관규칙은 지지도(support)와 신뢰도(confidence)라는 두 가지 척도에 의해 규칙을 찾아낸다. 지지도는 두 항목의 동시 발생이 얼마나 빈번했는가를 나타내며 신뢰도는 항목 A가 발생하는 경우 항목 B가 얼마나 발생하였는가를 나타낸다. 따라서 연관규칙 분석에 의해 발견된 규칙도 의미 없는 것일 가능성이 있다. 이는 연관규칙에 의해 발견된 규칙을 논리적으로 해석해야 할 필요가 있다는 것으로 시사하며 발견된 규칙 가운데 의미 있는 규칙을 선별하여 이용해야 한다는 것을 의미한다.

본 연구에서는 모든 변수를 이용하여 연관분석을 실시하였는데, 변수는 모두 130개이고 관측치는 모두 316개이다. 연관분석 결과 지지도와 신뢰도가 모두 100%인 관계가 360개가 확인이 되었다. 이 가운데 유사한 변수를 제거하고 모두 17개의 변수간 연간관계를 선택하였다. 물론, 최종 관계를 선택하는 데는 이론적 근거가 있는 것은 아니며, 이에는 과거의 경험 등 다소의 자의적인 기준과 변수간의 관계에 대한 논리적 근거 등이 동시에 반영될 수 있다.

한편 이렇게 선택된 관계를 변수로 전환하여 의사결정나무 모형의 투입요소로 이용하는 의사결정나무분석을 실시해 보았다. 우선 연관규칙에서 나온 관계를 바탕으로 형성된 변수만을 이용하여 의사결정나무 모형을 설정하였다. 그 결과, 앞에서의 여러 분석과 마찬가지로 수익률이 이상매매 선별에 중요한 요소로 작용하고 있으며 매수주문수량/매도주문수량 및 취소정정량/전체주문량 (각각 10일 평균)도 의미 있는 변수로 선택되었고, 이상매매를 이상매매로 분류하는 경우가 상당히 높게 나타났다. 그러나 이 모형의 성과를 판단하기 위해서는 보다 광범위한 자료를 이용한 검증이 필요하다.

4.5 거래소 통계적 적출모형과의 비교

본 절에서는 모형의 적합성을 검증하기 위해 데이터마이닝 기법에 의한 판정과 현재 증권거래소에서 이용하고 있는 통계적 적출 모형과 비교한다. 특히, 거래소 통계적 모형이 적출해내지 못한 종목 가운데 몇 개 종목을 데이터마이닝 기법이 이상매매로 분류해내는가를 보기로 한다. 그런데, 여기서 주의할 점은 데이터마이닝을 이용한 분석에서 전체 데이터를 훈련용 데이터와 검증용 데이터로 나누어 훈련용 데이터로 모형을 훈련시킨 후 검증용 데이터로 훈련된 모형을 테스트하였다는 것이다. 따라서 훈련용 데이터를 모형에 적용시킬 경우 그 성과는 당연히 높게 나타날 수밖에 없다. 앞의 사례에서 보면 이상매매 종목을 이상매매로 제대로 분류해낼 확률이 훈련용 데이터에서는 90% 수준으로 나타났으나 검증용 데이터에서는 60% 수준으로 낮아지는 경우가 있었는데, 이는 대부분의 데이터마이닝에서 나타나는 현상이다.

여기서는 훈련용 데이터와 검증용 데이터에 포함된 116개의 모든 이상매매 종목을 비교 대상으로 한다. 이는 데이터마이닝 프로그램이 훈련용 데이터와 검증용 데이터를 무작위로 나누어 추정 및 검증을 하기 때문이다. 한편, 대부분의 모형에서 훈련용 데이터와 검증용 데이터의 비율이 2:1이 되도록 설정하였기 때문에 본 절에서의 비교 결과는 과장 정도가 보다 클 것으로 예상된다.

두 번째로 주의할 점은 여기서의 분석에서는 앞에서와 마찬가지로 제2종오류(type 2 error, 즉, 정

상 종목을 이상매매 종목으로 잘 못 분류하는 오류)는 고려하지 않고 있다는 점이다. 이상매매 적출에서 중요한 것은 정상 종목이 이상매매 종목으로 분류되는 경우(제2종오류)가 다소 있더라도 이상매매 종목을 정상 종목으로 분류하는 것(제1종오류)을 최소화하는 것이다. 따라서 본 장에서 데이터마이닝 모형을 평가하는데 있어 제1종오류만을 고려하였다. 그러나 제2종오류가 커지는 경우에는 최종적으로 이상매매 여부를 판단하는데 보다 많은 비용과 노력이 필요하게 된다. 이 같은 점을 고려할 때, 두 모형의 성과를 정확히 비교하기 위해서는 제2종오류에 대한 분석이 필요하다. 거래소의 통계적 적출 모형의 경우, 제2종오류가 데이터마이닝 모형보다 크지 않을 것으로 추정된다. 따라서 제2종오류를 고려하는 경우 데이터마이닝 모형의 실제 성과는 본 분석에서의 성과보다 낮을 것으로 예상된다.

본 절의 비교에서 사용된 이상매매 종목은 모두 116개이나 모형에 따라 결손값을 제외하는 경우에는 관측치의 개수가 달라질 수 있다. 앞에서의 데이터마이닝 모형 분석 결과에 의하면, 모형 및 자료에 따라 다르게 나타나기는 하지만 가장 성과가 좋은 경우 이상매매를 이상매매로 분류할 가능성이 60% 수준으로 나타났다. 이러한 수치는 거래소의 통계적 적출 모형보다 다소 높은 수준이다. 그러나 앞에서 언급한 바와 같이, 제2종오류가 반영되어 있지 않기 때문에 데이터마이닝 모형이 유의하게 성과가 좋다고 평가하기는 곤란하다.

4.5.1 의사결정나무 모형 : 모든 변수를 적용한 경우

이 모형을 선택한 것은 앞에서의 여러 모형 가운데 이 모형의 성과가 비교적 좋은 편이기 때문이다. 물론, 여러 차례 지적한 바와 같이 관측치의 수가 충분하지 않기 때문에 검증용 데이터에서 나타난 성과가 향후에도 지속되지 않을 가능성도 있다. 이 모형을 설정하는데 일부 변수에 결손치가 있는 관측치는 제외하였기 때문에 여기서도 모형 설정에 제외되었던 관측치는 제외하고 분석한다. 즉, 일부 변수에 결손치가 있는 관측치를 제외한 98개 관측치를 대상으로 하였다. 의사결정나무에 의한 경우, 총 98개 관측치 가운데 이상매매로 판정된 종목의 비율은 67% 수준이다.

한편, 거래소의 통계적 적출 모형은 본 절에서 이용된 98개 관측치 가운데 이상매매 종목으로 판정한 비율은 50% 수준을 하회하였다. 한편, 여기서 눈에 띄는 사실은 의사결정나무 모형이 정상으로 잘 못 판정한 종목과 거래소 모형이 정상으로 잘 못 판정한 종목이 상당 부분 겹치고 있다는 것이다. 즉, 의사결정나무 모형이 제대로 분류해내지 못한 종목이 대부분 거래소 모형에 의해서도 제대로 선별되지 못한 종목이라는 것이다. 이는 거래소의 통계적 적출 모형과 의사결정나무가 서로 유사한 의사결정기준을 사용하고 있을 것이라는 것을 암시한다.

4.5.2 로짓 모형 : 모든 변수를 이용하는 경우

이 모형은 단계식 회귀분석에 의해 선택된 변수만을 이용하여 로짓분석을 실시한 것이다. 여기서도 변수에 결손치가 다수 있는 관측치를 제외하였기 때문에 이상매매 종목은 모두 98개이다. 이 자료에 대해서 거래소의 통계적 적출 모형의 이상매매 종목 분류 성공률은 앞에서보다 다소 높은 50%였다. 한편, 로짓 모형의 이상매매 적출 성공률은 71% 정도로 상대적으로 높게 나왔다.

로짓 모형은, 훈련용 데이터를 포함하여 볼 때,

거래소의 통계적 적출 모형이 이상매매로 제대로 분류하지 못한 종목 가운데 상당 부분을 이상매매 종목으로 분류하였다. 반면, 거래소 모형은 로짓 모형이 제대로 선별해내지 못한 종목 가운데 상당 부분을 이상매매 종목으로 분류하였다. 이는 의사결정나무에서와는 달리 로짓 모형과 거래소 모형은 서로 유사한 의사결정 기준을 사용하지 않고 있다는 것을 시사한다.

5. 결론 및 시사점

지금까지 증권거래소의 이상매매를 적출하기 위한 여러 데이터마이닝 방법들을 살펴보았다. 보다 효과적인 모형 구축을 위해 변수를 크게 일일자료 관련 변수와 일중자료 관련 변수로 나누어 다양한 조합을 시도하였다. 그 결과 실증적으로 우수한 성과를 보인 것인 로짓 모형이었으며 의사결정나무 모형도 보다 나은 성과를 보였다. 이상매매 적출은 본질적으로 주가 예측을 모형에 포함하고 있다. 그러나 주가 예측을 위해 이제까지 다양한 데이터마이닝 기법을 포함한 수많은 실무적 및 학술적 연구가 실시되어 왔으나 일관되게 우월한 성과를 보이는 모형은 없었다. 이러한 점이 바로 우리나라에서 통계적 시스템에 의해 이상매매를 적출해내는 것을 어렵게 하는 요소이다. 즉, 데이터마이닝은 결국 과거의 자료를 학습하여 동일한 패턴을 보이는 종목을 찾아내는 것인데 아무리 시세조종 종목이라 하더라도 주가 움직임의 패턴이 서로 같지는 않다는 것이다. 미국의 ADS는 데이터마이닝을 이용하여 시장조성자의 불공정 거래행위를 효과적으로 찾아내고 있는 것으로 알려지고 있다. 이처럼 ADS가 성공적으로 운영되는 것은 우선 그 대상이 시장조성자라는 이미 알려져 있는 제한된 집단이기 때문이다. 다음으로 이들의 불공정 행위 양태는 이미 알려져 있는 것으로 동일한 패턴을 보이기 때문이다. 이에 비해 우리나라에서의 이상매매 종목 적출은 알려지지 않은 미지의 투자자를 대상으로 알려지지 않은 양태의 거래행위를 적출해내는 것이므로 훈련에 의한 패턴 인식에 탁월한 성과를 보이는 데이터마이닝 기법들이 큰 효과를 낼 수 없는 것이다.

그러나 이러한 문제점에도 불구하고 본 연구에서 다룬 모형 가운데 일부는 상당한 성과를 보였으며, 추후 데이터의 누적과 모형에 대한 세부적 조정이 지속된다면 그 효과도 더 커질 것으로 예상된다. 본 연구에서는 지도학습 기법으로 인공신경망, 로짓, 그리고 의사결정나무를 사용하였다. 이 가운데 인공신경망 모형은 성과가 좋은 편이 아니었는데, 이는 특히 관측치의 부족에 기인한다. 이상매매 적출에 인공신경망을 적용하기 위해서는 상당 큰 규모의 관측치가 필요한데 현실적으로 그 수준의 데이터를 수집하기는 어려울 것으로 보인다.

로짓 모형의 경우는 추정된 계수의 일관성을 위해서는 독립변수의 수에 비례하여 관측치의 수가 늘어나야 한다. 그러나 관측치의 수가 제한적이므로 부득이 독립변수의 수를 축소해야 한다. 본 보고서에서는 단계적 회귀분석을 통해 변수를 선택하였는데 변수 3개로 구성된 비교적 간단한 모형이 특히 성과가 높은 것으로 나타났다. 한편, 의사결정나무의 경우는 자체적으로 변수를 축소하는 메커니즘이 있으므로 비교적 변수의 수에 대한 제약이 적다. 대부분의 경우, 그 성과가 상대적으로 좋게 나타났다.

이러한 점을 고려할 때, 현재 적용 가능한 모형은 로짓 모형과 의사결정나무인 것으로 판단된다. 물론, 현재 사용 가능한 데이터의 수를 고려할 때

이러한 모형이 안정적이지는 못한 것으로 보인다. 따라서 일단 모형을 구축하더라도 모형을 개선하는 노력이 지속되어야 할 것으로 보인다. 한편, 연관분석을 통해 얻은 관계를 변수로 반영한 모형의 성과도 좋은 것으로 나타났으나 이는 좀더 검증되어야할 모형으로 생각된다.

마지막으로 도출된 실증결과에 비추어 데이터마이닝을 이용한 이상매매 적출모형에 대한 구체적 검토 및 활용 여부의 문제와 데이터마이닝 적출모형을 활용하는 경우 현행 통계적 모형과의 관계 정립에 관한 문제에 대한 시사점들을 살펴보면 다음과 같다.

첫째, 데이터마이닝 이상매매 적출모형은 최적 적출모형의 구축 등과 같은 구체적 사안에 대해 일정기간 동안의 추가적 검토와 테스트를 거쳐 긍정적 결과가 나오는 경우 실무적 활용을 고려해 볼 수 있을 것으로 판단된다. 그 중 특히 연관규칙과 연계된 의사결정나무 모형과 로짓 모형은 기존의 통계적 모형에 의해 적출되지 않았지만 풍문이나 금융감독원의 의뢰를 통해 밝혀진 시세조종 종목들의 일부를 찾아냈을 뿐만 아니라 실제 시세조종 종목 자료들을 이용해서 측정한 적출효율성 측면에서 우수한 성과를 보이기도 하였다. 그러나 본 연구에서 제시된 데이터마이닝 적출모형은 소수의 후보군을 선정하여 얻어진 것으로서 검토 가능한 전체 모형으로부터 도출된 최적 모형이 아닐 가능성이 있으며, 또한 모형에 투입되는 데이터와 변수들의 선정에 따라 모형의 적출성과가 민감하게 변하는 문제점이 있기 때문에 본 연구에서 사용된 모형을 실무에 당장 활용하기에는 무리가 있다고 본다.

둘째는, 데이터마이닝 모형을 도입할 경우 현재의 통계적 모형과 데이터마이닝 모형에 대한 활용방안에 관한 문제로서, 본 연구는 데이터마이닝 모형이 도입되더라도 상당기간 동안 통계적 모형을 이상매매 적출을 위한 주된 모형으로 사용하고, 데이터마이닝 모형은 주로 풍문, 공시 등에 의해서 개시된 시세조종 혐의기업들의 적출과 심층적인 심리작업에서 시세조종 종목들의 과거 특성에 비추어 관심종목이 유사한 특성을 갖고 있는지 분석하는 과정에서 검토변수들을 주기적(분기별 혹은 반기별)으로 선정하고 그들의 우선순위를 설정하는데 활용하는 것이 바람직하다고 여겨진다. 즉, 데이터마이닝 모형을 기존의 통계적 적출모형을 대체하는 수단으로 활용하기 보다는 이상매매의 적출효율성 및 심리업무의 정확성을 높이기 위한 보조수단의 역할을 담당하게 하는 것이 거래소의 이상매매 적출 및 심리업무에 도움을 줄 수 있는 방법이라 생각된다.

참고문헌

[1] 강현철 · 한상태 · 최종후 · 김은석 · 김미경, *데이터마이닝 : 방법론 및 활용*, 자유아카데미, 2002.

[2] 김광용, "데이터마이닝 기법의 성과평가 및 새로운 위험분류측정에 관한 실증적 연구", *보험개발연구*, 12권 2호(2001), 133-166.

[3] 김동조, *주식작전 大해부*, 마이웨이라이프, 2003.

[4] 용환승 역(Addriaan and Zantinge), *데이터마이닝*, 그린, 1998.

[5] 이영섭 역(Olivia Parr Rud), *Data Mining Cookbook*, 교우사, 2003.

[6] 최종후 · 한상태 · 강현철 · 김은석 · 김미경 · 이성건, *데이터마이닝 : 기능과 사용법*, 자유아카데미, 2001.

[7] Bell, Timothy B., "Neural Nets or the Logit

Model? A Comparison of Each Model's Ability to Predict Commercial Bank Failures", Intellignet Systems in Accounting, Finance and Management, Vol.6(1997), 249-264.

[8] Kim, Steven H., *Data Mining in Finance Sigma*, Consulting Group, 1999.

[9] Kirkland, J. Dale , Ted E. Senator, James J. Hayden, Tomasz Dybala, Henry G. Goldberg and Ping Shryr, "The NASD Regulation Advanced-Detection System", *AI Magazine*, Spring 1999, Vol.20 i1, 55-64.

[10] Roiger, R. R., and Michael W. Geatz, *Data Mining: A Tutorial-Based Primer*, Addison Wesley, 2003.

[11] SAS Institutue Inc., *Applying Data Mining Techniques Using Enterprise Miner: Course Notes*, 2002.

[12] Senator, Ted E., "Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation's Advanced Detection System (ADS)", *KDD-2000*, ACM 2000, 44-54.

Abstract

Detection of Stock Price Manipulation : A Data Mining Approach

Chung-Hun Hong* · Sung Mahn Ahn* · Kyung Woo Wee**

In this paper, we discuss a data mining approach to detection of stock price manipulation in the Korean stock market. First of all, we review current methods which is being exercised in the Korean stock market as well as in the US stock market. And then we apply data mining techniques to the problem using data from the Korean stock market and discuss the results along with their implications.

Key words : Data Mining, Stock Price Manipulation, Korean Stock Market

* School of Business Administration, Kookmin University
** Department of Business Administration, Sookmyung Women's University