Journal of the Korean Data & Information Science Society
2013, **24**(2), 223–233

http://dx.doi.org/10.7465/jkdi.2013.24.2.223
한국데이터정보과학회지

외국인 투자자의 비정상적 중·장기매도성향패턴예측을 위한 지능형 조기경보시스템 구축[†]

오경주[1] · 김영민[2]

[12]연세대학교 정보산업공학과

접수 2012년 11월 20일, 수정 2012년 12월 11일, 게재확정 2013년 1월 28일

요 약

본 연구는 외국인 투자자의 대량매도구간을 서포트 벡터 머신 알고리즘을 통해 모형을 구축하여 발생 가능한 대량매도기간을 사전에 방지할 수 있는 지능형 조기경보시스템을 구축하였다. 이러한 방법은 기존의 Son 등 (2009), Ahn 등 (2011)이 제시한 방법을 토대로 지능형 조기경보시스템에 대한 예측성과를 개선시켰으며, 더 나아가 최근까지 예측성과를 살펴봄으로써 조기경보시스템의 역할을 수행할 수 있는지를 살펴보았다. 또한 구축된 EWSFI는 국내주식시장뿐만 아니라 환율 및 원유시장 등 다양한 경제 분야에서 활용될 수 있는 가능성을 시사하고 있으며, 시장상황의 위기를 사전에 예측하여 예상되는 충격을 줄일 수 있을 것이다.

주요용어: 외국인 투자자, 장기매도성향패턴, 지능형 조기경보시스템.

1. 서론

1990년대 초반부터 시작된 한국, 홍콩, 싱가포르, 타이완 등의 아시아 신흥시장은 세계 금융시장에 편입되었다 (Ghysels 등, 2005). 정보통신 발달에 따른 금융시장의 세계화와 규제완화로 인해 외국인 투자자 (foreign investors)은 아시아 신흥시장에서 중요한 투자 집단으로 등장하였으며, 심각한 내·외부적 충격이 발생할 경우에 외국인 투자자의 투자 판단에 따라 해당 시장의 방향이 결정되었다. 2004년 4월까지 한국 주식 시장에서 시가 총액의 거의 절반을 소유하고 있었던 외국인 투자자는 1997년 아시아 경제 위기, 1998년 러시아 모라토리엄 선언, 1999년 대우 그룹 유동성 위기, 2000년 현대그룹 유동성 위기 및 2001년 9/11 테러와 같은 국내외 증시에 부정적 영향을 미치는 사건이 발생할 때마다 자신들이 보유하고 있던 주식을 대량매도함으로써 국내 주식시장의 20∼40% 정도의 폭락 장세를 지속적으로 주도하였다 (Kim 등, 1999; Choe 등, 1999; Kim과 Kwon, 2003). 실제로 외국인 투자자들의 대량매도로 인한 주가하락의 가능성에 관한 실증연구가 발표되고 있으며, Radelet와 Sachs (1998), Kim과 Wei (1999), Choe 등 (1999)은 1997년 한국금융시장의 위기는 외국인 투자자의 대량 매도로 인해 더욱 가속화 되었다고 언급하였다.

또한, 2008년에는 미국의 대표적 투자은행인 리먼브라더스의 파산 신청, 메릴린치 매각 등 일련의 대형금융 사건과 동시에 찾아온 국내 주식시장에서의 외국인 투자자의 대량 매도는 주가지수의 급락과 함

[†] 이 논문은 2010년 LG연암문화재단에서 지원하는 해외연구교수지원금을 받아 수행되었음.
[1] 교신저자 : (120-749) 서울특별시 서대문구 신촌동 134번지, 연세대학교 정보산업공학과, 부교수.
E-mail: johanoh@yonsei.ac.kr
[2] (120-749) 서울특별시 서대문구 신촌동 134번지, 연세대학교 정보산업공학과, 박사과정.

께 많은 투자자들에게 엄청난 피해를 안겨주었다. 이러한 이유로 미래에 갑작스럽게 찾아오는 국내·외 금융위기에 기인하는 국내주식시장의 급락을 효율적으로 제어하기 위하여 외국인 투자자의 비정상적인 대량매도성향을 미리 탐지하거나 예측할 수 있는 지능형 조기경보시스템의 개발이 필요하다.

기존 금융위기 조기경보시스템에 대한 연구를 살펴보면, 많은 학자들이 통계적 모형들과 시계열 모형들을 이용하여 금융위기 탐지를 위한 조기경보시스템을 개발하였다 (Eichengreen 등, 1996; Frankel와 Rose, 1996; Kaminsky 등, 1998; Kaminsky와 Reinhart, 1999; Goldstein 등, 2000; Edison, 2000). 또한 경제 위기는 해당 국가의 취약한 기초경제조건 (fundamental status)에 기인한다는 가정 하에 모형이 개발되었으며 (Krugman, 1979; Obstfeld, 1986; Eichengree 등, 1995), 향후 1∼2년 이내에 위기가 올 것인가, 아닌가를 예측하였다.

본 연구에서는 외국인 투자자의 순매수금액을 바탕으로 시장상황을 정립하며, 앞으로 발생 가능한 외국인 투자자의 대량매도패턴 움직임을 사전에 방지 할 수 있는 조기경보시스템 (early warning system for foreign investors; EWSFI) 구축하고자 한다. 이러한 방법은 Ahn 등 (2011)이 제안한 Lag-ℓ 예측방법과 기계학습 알고리즘들 (machine learning algorithms)을 이용하여 구축된 지능형 조기경보시스템을 바탕으로 중·장기예측에 대한 성과를 개선하고, 더 나아가 현재시점까지 구축된 EWSFI의 성과를 확인하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 EWSFI를 구축하는 단계 및 학습 알고리즘에 대하여 설명하며, 3절에서는 제안 시스템의 실증분석 결과를 토대로 시스템의 유용성을 분석하였다. 마지막으로 결론부분에서는 본 연구의 기대효과 및 향후 연구에 대해 서술하였다.

2. 연구방법

2.1. EWSFI 구축

EWSFI 구축 과정은 크게 2단계로 구분할 수 있다. 1단계는 외국인 투자자의 주식 순매수 금액을 바탕으로 미래의 주식시장 상황을 안정 (stable period; SP), 불안정 (transition period; TP), 위기 (crisis period; CP)구간으로 판단하는 오라클 분류기 (oracle classifier, Oc)를 구축하며, 2단계는 기계학습 알고리즘인 서포트 벡터 머신 (support vector machine; SVM)알고리즘을 이용한 훈련 분류기 (trained classifier; Tc)를 구축한 후, 미래의 주식시장에 대한 상황을 예측한다.

1단계. 오라클 분류기 구축

오라클 분류기를 구축하는 데 있어서 첫 번째 단계는 외국인 투자자의 순매수금액을 기준 삼아 주식시장의 상황을 SP, TP, CP 구간으로 정의한다 (Son 등, 2009; Ahn 등, 2011). 본 연구에서도 CP 구간을 외국인 투자자가 비상계획에 따라 대규모로 주식을 매도하는 시기로 정의하며, TP 구간은 비상계획의 초기 비상계획 시 외국인 투자자가 주식 순매수 기조에서 순매도 기조로 전환하는 기간으로 정의한다. SP 구간의 경우에는 TP와 CP 구간이 아닌 구간으로 정의한다. 앞의 정의에 기반을 두고 SP, TP 및 CP는 외국인 투자자의 일별, 주별, 월별, 분기별 누적 순매수포지션 (net cumulative buying position)에 따라 구체적 수치로 정의되며, 3절 Table 3.1에 제시되어 있다. 오라클 분류기는 다음 식 (2.1)으로 표현 할 수 있다.

$$Oc : OX \rightarrow OY \tag{2.1}$$

단, OY(=1,2,3)는 여러 가지 기준의 O에 의해 결정되며, 각각 SP, TP 및 CP에 대응되는 값이다. OX는 O_{x1}, O_{x2}, O_{x3}, O_{x4}로써 외국인 투자자의 일별, 주별, 월별, 분기별 누적 순매수 포지션을 의미하며, 일별은 1일, 주별은 5일, 월별은 20일, 분기별은 60일로 정의한다.

2단계. Lag ℓ 분류기 구축

일단 Oc가 성공적으로 구축된 후에, 다음과 같은 절차에 따라 Lag ℓ 분류기를 구축한다. Lag ℓ 분류기를 구축하기 위해서는 변수선택과정을 거쳐 외국인 투자자의 움직임을 가장 잘 설명하는 입력변수들 $(X_1, X_2, X_3, \cdots, X_p)$과 오라클 분류기를 통해 생성된 시장상황인 출력변수 값을 Y라 설정한다. 양의 정수 ℓ에 대한 학습 데이터셋 (training dataset)은 다음 식으로 표현된다.

$$(X_{11}, X_{21}, \cdots, X_{p1}, Y_{1+\ell}), \cdots, (X_{1n}, X_{2n}, \cdots, X_{pn}, Y_{n+\ell}) \tag{2.2}$$

단, $Y_{1+\ell} = OY_{1+\ell}, \cdots, Y_{n+\ell} = OY_{n+\ell}$ 은 식 (2.1)의 Oc에서 도출된 값이다. 다음으로 식 (2.2)는 학습 데이터셋을 기반으로, 서포트 벡터 머신 알고리즘을 이용하여, Lag ℓ 분류기를 구축한다.

$$f_\ell : X \to Y \tag{2.3}$$

식 (2.3)의 f_ℓ은 서포트 벡터 머신 알고리즘을 통해 학습된 Lag ℓ의 예측 모형이다. 한편 f_ℓ의 정확도는 적절한 학습 데이터셋을 선택하는 것에 따라 테스트 데이터셋의 예측성과를 결정하기 때문에, 적합한 학습 데이터셋을 추출하는 것이 중요한 과정중 하나이다. 따라서 과거의 외국인 투자자에 의해 발생한 대량매도구간을 파악하여 학습 데이터셋을 추출하는 것이 적합한 기간이라 할 수 있다. 대량매도구간은 각각의 일별, 주별, 월별, 분기별 외국인 투자자의 누적 순매수 금액을 분석함으로써 발견할 수 있으며, 구간의 크기 n은 식 (2.2)의 학습 데이터셋으로 설정할 수 있다. 또한 입력 변수 $(X_1, X_2, X_3, \cdots, X_p)$는 외국인 투자자의 움직임을 가장 잘 반영할 수 있는 변수들을 선택하여야 한다. 이러한 변수들의 선택은 EWSFI의 유용성 제고를 위한 중요한 과정 중 하나이기 때문에, 다양한 변수 변환 (transformation)을 통해 파생변수를 생성하여야 하며, 전문가의 의견도 변수 선택 과정에 고려되어야 한다. 본 연구에서는 Son 등 (2009)과 Ahn 등 (2011)이 제시한 변수들을 선택하였으며, 이는 3절 실증분석에 자세히 설명되어 있다. 마지막으로 입력변수의 값이 극단적으로 크거나, 작은 값을 가지는 입력변수들은 f_ℓ을 학습하는데 영향을 주지 않게 하기 위해 입력 변수들을 [0,1]로 선형 변환한다. 실제로 특정구간으로 입력변수를 변환하는 것이 $(X_{1t}, \cdots, X_{pn})$ 과 Y_{t+l} 사이의 f_ℓ를 정상상태 (stationary condition)로 만들기 때문에, 예측오차를 줄이는 데 중요한 역할을 한다 (Lapedes와 Farber, 1988).

2.2. 서포트 벡터 머신

서포트 벡터 머신 (support vector machine; SVM) 알고리즘은 Vapnik에 의해 개발된 분류기법으로, 최근 데이터 마이닝과 패턴인식 분야 등에 널리 사용되고 있다. 기존의 기계학습 알고리즘들은 경험적 위험 최소화 (empirical risk minimization) 방법에 기초하고 있는 반면에, 서포트 벡터 머신의 경우에는 각 군집 사이의 여백을 최대화하는 최적 분류 초평면 (optimal separating hyperplane; OSH)을 통한 구조적 위험 최소화 (structural risk minimization) 방법에 기초하고 있다. 따라서 과적합 (overfitting)을 피할 수 있으며 볼록함수 (convex function)를 최소화하는 방법으로 학습을 진행하기 때문에 전역 최적해 (global optimal solution)를 찾을 수 있어 다른 기계학습 알고리즘들보다 우수한 기법으로 알려져 있다 (Burges, 1998). 특히, 서포트 벡터 머신의 장점은 서포트 벡터 (support vector)라고 불리는 소수의 데이터만을 최종적으로 학습에 사용하기 때문에, 일반적으로 적은 양의 학습 데이터로도 우수한 예측성과를 나타낸다. 본 연구에서는 서포트 벡터 머신의 실험용 소프트웨어인 LIBSVM (http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/)을 이용하였다.

3. 실증분석

실증분석에서 사용된 자료는 코스콤에서 제공하는 프로그램 CHECKExpert로부터 구하였으며, 1999.1~2012.12까지 자료를 수집하였다. 본 연구는 기존의 Ahn 등 (2011)이 제시한 방법을 사용하

여 동일 기간 (1999.1∼2004.4)내에서 서포트 벡터 머신 알고리즘으로 구축된 EWSFI의 유용성을 살펴보고, 더 나아가 2008년 글로벌 금융위기를 포함한 2012.12까지 EWSFI의 성과를 살펴보고자 한다.

EWSFI의 첫 번째 단계는 오라클 분류기를 통하여 시장상황을 정의하는 것이다. 즉, 외국인 투자자의 코스피 일별 순매수금액 (NPG)을 바탕으로 주별, 월별, 분기별로 누적시켜, 제시된 Table 3.2의 오라클 분류기의 분류 기준에 따라 국내주식시장 상황을 일별로 안정 (SP), 불안정 (TP), 위기 (CP) 구간으로 정의한다. 여기서 주별 (NPG_5), 월별 (NPG_20), 분기별 (NPG_60)은 외국인 투자자의 일별 순매수금액을 누적시켜 이동 평균 (주별은 5일, 월별은 20일, 분기별은 60일)된 값을 의미하며, 이에 대한 기초통계량은 Table 3.1에 제시되어 있다.

Table 3.1 Descriptive statistic analysis of NPG

variable	NPG	NPG_5	NPG_20	NPG_60
mean	9,538	47,696	181,458	539,027
max	1,719,998	4,089,045	8,266,912	15,171,766
min	-1,309,443	-3,965,646	-9,611,295	-17,426,896
Standard deviation	228,672	804,545	2,370,716	5,494,621

Table 3.2 Oracle classification rule for market condition

	Classification rule
$f_0(1)$	If NPG_60 < -2.4 or NPG_20 < -1.6 or NPG_5 < -0.8 or NPG < -0.4then Y = 3 (CP) else If NPG_60 < -1.2 or NPG_20 < -0.8 or NPG_5 < -0.4 or NPG < -0.15 then Y = 2 (TP) else Y = 1 (SP)
$f_0(2)$	If NPG_60 < -3.0 or NPG_20 < -2.0 or NPG_5 < -1.0 or NPG < -0.5 then Y = 3 (CP) else If NPG_60 < -1.5 or NPG_20 < -1.0 or NPG_5 < -0.5 or NPG < -0.2 then Y = 2 (TP) else Y = 1 (SP)
$f_0(3)$	If NPG_60 < -4.0 or NPG_20 < -3.0 or NPG_5 < -1.5 or NPG < -0.7 then Y = 3 (CP) else If NPG_60 < -2.0 or NPG_20 < -1.5 or NPG_5 < -0.8 or NPG < -0.4then Y = 2 (TP) else Y = 1 (SP)

예를 들어, 오라클 분류기준 중 $f_0(2)$의 조건을 살펴보면, 누적 순매수금액이 분기별 (-3,000억원), 월별 (-2,000억원), 주별 (-1,000억원), 일별 (-500억원) 보다 작다는 조건을 하나만 만족하게 되면 시장상황을 CP라고 판단한다. TP의 경우에는 누적 순매수 금액이 분기별 (-1,500억원), 월별 (-1,000억), 주별 (-500억), 일별 (-200억) 보다 작다는 하나의 조건을 만족한 경우이다. SP 경우에는 CP와 TP가 아닌 경우이다. 여기에서는 $f_0(2)$가 기준이 되며, $f_0(1)$과 $f_0(3)$은 분류기준의 비교 목적으로 사용된다. 그 이유는 오라클 분류기준은 시장의 변화에 따라 조정 가능하기 때문이다. 즉, $f_0(1)$은 $f_0(2)$ 보다 CP 상태를 더 많이 판단하기 때문에 민감한 분류기라고 할 수 있는 반면에, $f_0(3)$의 경우에는 $f_0(2)$보다 CP 상태를 적게 판단하기 때문에 보수적인 분류기라고 할 수 있다. 첫 번째 실험에 사용된 총 데이터 구간은 1999.1-2004.4이며, 오라클 분류기 기준인 $f_0(1)$, $f_0(2)$, $f_0(3)$에 따라 시장상황을 판단 (SP, TP, CP)하는 출력변수 (Y)가 일별로 생성된다. 생성된 출력변수는 EWSFI의 2단계인 Lag ℓ 분류기에서 출력변수 (Y)로 정의된다.

EWSFI의 2단계인 Lag ℓ 분류기 구축 단계는 시장상황을 예측하는 모델을 구축하기 위하여 서포트 벡터 머신 알고리즘을 사용하였다. 기계학습 알고리즘들은 학습 데이터셋을 통하여 모형이 생성되기 때문에 적절한 학습 데이터셋의 선정이 중요하다. 이 데이터 셋에 해당하는 구간은 Ahn 등 (2011)이 제시한 대량매도구간 (enormous selling periods) 1999.7∼1999.9 (ESP99)과 2002.2∼2002.4 (ESP02)을 학습 구간으로 사용하였다. 그 이유는 본 연구의 목적 중 하나가 서포트 벡터 머신 알고리즘의 유용성을 확인하는 것이기 때문에 데이터셋을 동일 시 하였다. Ahn 등 (2011)이 제시한 비정상적 대량매도구간 선정은 외국인 투자자의 누적순매수금액 중 60일 이동 평균된 NPG_60 (분기별)을 확인하여 대량매도구간을 선정하였으며, 그 구간 전후로 오라클 분류기준에 의해 생성된 안정, 불안정, 위기 구간을 선정한다. 이와 같은 방법으로 선정된 구간은 Table 3.3과 Table 3.4에 제시되어 있으며, Lag ℓ 분류기에서 학습 데이터셋으로 사용된다. 즉, 해당 구간의 입력변수들과 오라클 분류기를 통해 생성된 시장상

황 (출력변수)이 서포트 벡터 머신을 통하여 학습된다. 그리고 f_{20}, f_{60}은 20일 후, 60일 후의 시장상황을 예측하는 것을 의미한다. 단, 2002년의 대량매도구간인 ESP02 경우에는 오라클 분류기준 중 하나인 $f_0(3)$ 기준을 만족시키는 데이터 구간이 존재하지 않아 $f_0(1)$과 $f_0(2)$만 사용하였다. 따라서 Lag ℓ 분류기를 서포트 벡터 머신 알고리즘으로 학습시키기 위한 학습 데이터구간은 총 10 구간이며, 총 데이터 기간 중 대량매도구간을 제외한 데이터셋은 모형의 성과를 측정하기 위한 테스트 데이터셋으로 사용된다.

Table 3.3 Training period for f_{20} (monthly)

Year	Rule	SP	TP	CP
1999	$f_0(1)$	99.05.21∼99.06.11	99.06.14∼99.07.05	99.07.06∼99.07.27
	$f_0(2)$	99.05.24∼99.06.14	99.06.15∼99.07.06	99.07.07∼99.07.28
	$f_0(3)$	99.05.27∼99.06.28	99.06.29∼99.08.06	99.08.09∼99.09.08
2002	$f_0(1)$	01.12.28∼02.01.21	02.01.22∼02.01.25	02.02.15∼02.02.15
		02.01.28∼02.02.14	02.02.18∼02.03.07	02.03.08∼02.03.18
			02.03.19∼02.03.29	02.04.01∼02.04.25
	$f_0(2)$	01.12.27∼02.01.23	02.01.24∼02.01.25	02.03.13∼02.03.15
		01.01.28∼02.02.08	02.02.14∼02.03.08	02.04.09∼02.05.14
			02.03.18∼02.04.08	

Table 3.4 Training period for f_{60} (quarterly)

Year	Rule	SP	TP	CP
1999	$f_0(1)$	99.03.24∼99.04.15	99.04.16∼99.05.10	99.05.11∼99.06.01
	$f_0(2)$	99.03.25∼99.04.16	99.04.19∼99.05.11	99.05.12∼99.06.02
	$f_0(3)$	99.03.30∼99.04.30	99.05.03∼99.06.03	99.06.04∼-99.07.06
2002	$f_0(1)$	01.11.01∼01.11.21	01.11.22∼01.11.27	01.12.13∼01.12.13
		01.11.28∼01.12.12	01.12.14∼02.01.04	02.01.07∼02.01.15
			02.01.16∼02.01.28	02.01.29∼02.02.26
	$f_0(2)$	01.10.31∼01.11.23	01.11.26∼01.11.27	02.01.04∼02.01.08
		01.11.28∼01.12.11	01.12.12∼02.01.03	02.01.25∼02.03.08
			02.01.09∼02.01.24	

입력변수들도 마찬가지로 Ahn 등 (2011)이 제시한 외국인 투자자의 순매수금액 (NPG), KOSPI 지수 (SPI), 환률 (FER), 다우존스지수 (DJI), 외국인 기관투자자의 KOSPI200 지수 선물 순매수금액 (NPI)이며, 각각의 원계열과 파생된 변수들을 사용하였다. 이러한 입력변수들은 총 18개로 Table 3.4에 제시되어 있으며, IND는 원계열, MA는 이동평균, MV는 이동분산을 의미하며 괄호안의 숫자는 n-일수를 의미한다. 즉, MA(10)은 10일 이동평균을 의미한다.

Table 3.5 List of input and output variable for EWSFI

Variable names	Input variables	Output variables
NPG	IND, MA(10), MA(20), MA(60), MV(20)	Y(1=SP, 2=TP, 3=CP)
SPI	IND, MA(10), MA(20), MV(20)	
FER	IND, MA(20), MV(20)	
DJI	IND, MA(20), MV(20)	
NPI	IND, MA(10), MV(20)	

본 연구에서 구축된 Lag ℓ 분류기는 오라클 분류기를 통해 생성된 출력변수 (Y)인 시장상황과 Table 3.5에 제시된 입력변수들이 서포트 벡터 머신 알고리즘를 통해 대량매도구간을 학습시킨다. 따라서 학습된 Lag ℓ 분류기는 외국인 투자자의 중·장기매도 패턴을 학습하여 미래의 월별, 분기별 시장 상황을 예측한다. 즉, 20일 후, 60일 후의 시장상황을 일별로 지속적으로 모니터닝 할 수 있도록 예측한다. 구축된 EWSFI를 통해 테스트 데이터셋에 대한 예측 성과는 Table 3.6에 제시되어 있으며, 1999과 2002년도는 학습 데이터셋을 의미하는 대량매도기간 ESP99, ESP02이며, 오라클 분류기준에 따라 각각의 성과를 살펴볼 수 있으며, f_{20} (monthly), f_{60} (quarterly)은 20일 후, 60일 후의 시장상황을 예측

하는 것을 의미한다. 예측성과는 1단계에서 구축한 오라클 분류기를 통한 시장 상황과 서포트 벡터 머신 알고리즘을 통하여 학습된 Lag ℓ 분류기를 통한 예측결과를 비교하여 정분류율을 계산하였다.

Table 3.6 Training and testing rates (%) of EWSFI trained by SVM

Year	Rule	Datasets	f_{20} (monthly)	f_{60} (quarterly)
1999	$f_0(1)$	training	91.67	100
		testing	72.89	52.47
	$f_0(2)$	training	97.92	97.92
		testing	80.38	51.73
	$f_0(3)$	training	81.16	98.65
		testing	89.65	60.24
2002	$f_0(1)$	training	69.23	74.36
		testing	86.36	45.61
	$f_0(2)$	training	76.19	100
		testing	86.46	56.36
summary in testing		Mean	83.14	53.28
		Standard deviation	6.64	5.47

서포트 벡터 머신 알고리즘으로 학습된 EWSFI의 예측성과를 살펴보면, 기존의 Ahn 등 (2011)이 사용한 알고리즘들 (multinomial logistic regression, decision tree, case-based reasoning, artificial neural network)의 결과를 비교해 본 결과,서포트 벡터 머신 알고리즘의 유용성을 확인할 수 있었다.

앞에서 구축된 EWSFI를 기반으로 분석 기간을 2012.12까지 확장하여 EWSFI가 2008년 금융위기를 포함한 기간내에서도 조기경보시스템으로써 유용한지 확인하고자 한다. 그러나 분석 기간이 길어짐으로써 외국인 투자자의 누적 순매수금액으로 시장상황을 판단하는 오라클 분류기의 기준값들이 적합하지 않다. 그 이유는 2008년의 코스피지수가 2004년 이전에 비해 두 배 이상 성장하였기 때문이다. 이러한 양적 성장으로 말미암아 2004년 이전 코스피 1%를 올리기 위해 필요한 금액은 2008년에는 두 배 이상 소요됨을 의미한다. 따라서 본 연구에서는 이러한 문제점을 해결하기 위하여 외국인 투자자의 순매수 금액을 코스피의 시가총액으로 나눈 비율 (NPGR)을 가지고 오라클 분류기준을 조정할 필요가 있다. 이를 확인하기 위하여 Figure 3.1과 Figure 3.2을 살펴보면 외국인 투자자의 순매수금액 자체보다는 코스피 시가총액으로 나눈 비율을 토대로 시장상황을 정의하는 오라클 분류기를 구축하는 것이 적합함을 알 수 있다.

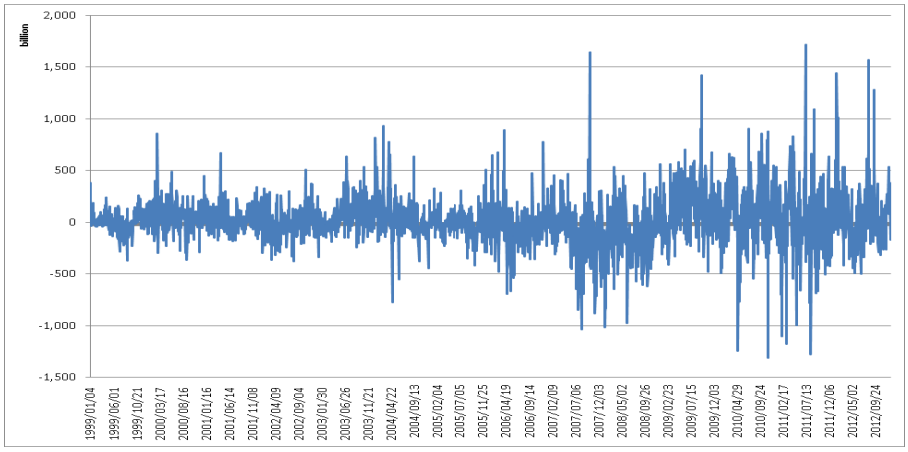

Figure 3.1 Daily NPG from January 1999 to December 2012

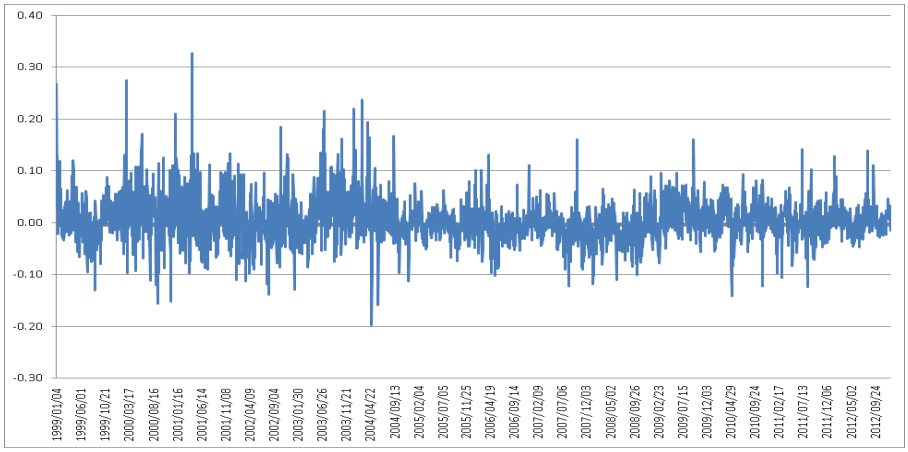

Figure 3.2 Daily NPGR from January 1999 to December 2012

따라서 시장 상황을 정의하는 오라클 분류기는 외국인 투자자의 순매수금액이 아닌 코스피 시가총액으로 나뉜 NPGR을 사용한다. 앞에서 제시된 방법과 동일하게 누적비율을 사용하여 일별, 주별, 월별, 분기별에 따라 시장상황을 정의하는 것은 동일하다. 단, 기존 방법에서는 오라클 분류기가 위기 (CP)를 판단하는 기준에 따라 민감한 분류기, 보수적인 분류기를 구축하여 각각 예측성과를 측정하였지만, 이번 실험에서는 한 가지 기준으로만 시장 상황을 판단하였다. 분석에 사용된 총 데이터는 1999.1-2012.12까지이며, 오라클 분류기를 통한 시장상황을 안정 (SP=1), 불안정 (TP=2), 위기 (CP=3) 3가지로 나뉘진 경우와 안정 (SP=1)와 위기 (CP=3)로 분류되는 오라클 분류기를 구축한 후, 서포트 벡터 머신으로 학습된 Lag ℓ를 통하여 향후 20일 후의 시장상황을 일별로 지속적으로 시장상황을 예측하였다.

Figure 3.3과 Figure 3.4는 오라클 분류기를 통한 시장상황과 코스피지수의 흐름을 보여주고 있으며, 2가지 모두 코스피지수의 등락에 따라 시장상황을 잘 나타내 주고 있음을 확인할 수 있다. 특히, 2007년의 서브프라임 모기지 (subprime mortgage)사태와 2008년도의 글로벌 금융위기 부분을 위기 (CP)와 불안정 (TP)구간으로 분류하는 것을 알 수 있다. 이와 같은 결과는 외국인 투자자의 순매도금액을 시가총액으로 나눈 비율 (NPGR) 바탕으로 이루어진 오라클 분류기가 시장상황을 잘 반영하고 있는 것을 알 수 있다.

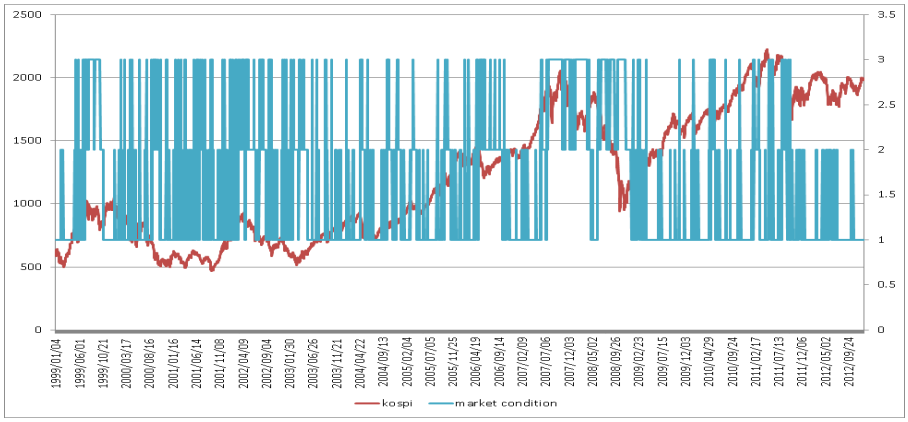

Figure 3.3 Classification result of oracle classifier (1:SP, 2:TP, 3:CP)

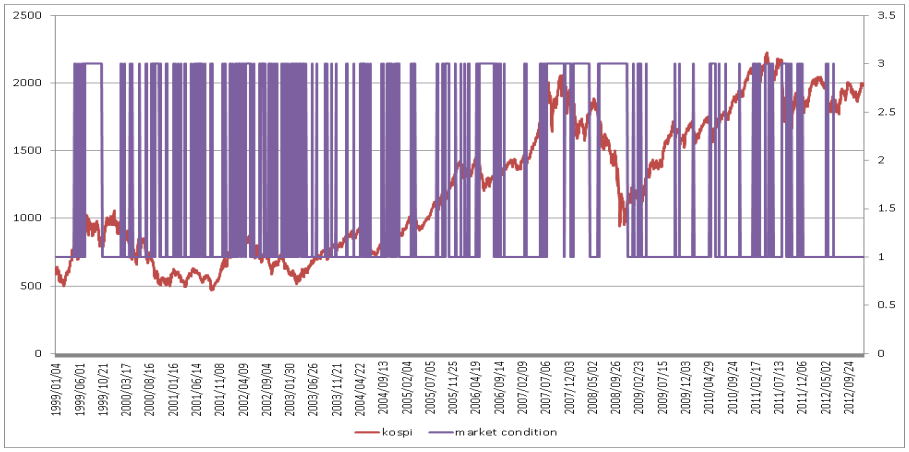

Figure 3.4 Classification result of oracle classifier (1:SP, 3:CP)

또한, 오라클 분류기준을 NPGR으로 조정한 후, EWSFI의 2단계인 Lag ℓ 분류기를 구축하기 위하여 학습 데이터셋이 필요하다. 즉, 외국인 투자자의 대량매도구간을 파악하여 그 해당 구간을 서포트 벡터 머신 알고리즘을 통해 Lag ℓ 분류기를 구축한다. 이러한 외국인 투자자의 대량매도구간은 Figure 3.5에 제시된 NPGR 60일 누적 이동 평균을 살펴봄으로써 1999.7-1999.11, 2002.2-2002.11 구간이 대량매도구간임을 확인할 수 있었다. 단, 2007년과 2008년 역시 대량매도구간임을 확인할 수 있지만, 본 연구의 목적 중 하나가 금융위기를 포함한 구간에서도 EWSFI의 유용성을 파악하는 것이기 때문에 1999년과 2002년의 대량매도구간을 학습 데이터셋으로 선정하였으며, 서포트 벡터 머신 알고리즘을 활용하여 Lag ℓ 분류기를 구축하였다.

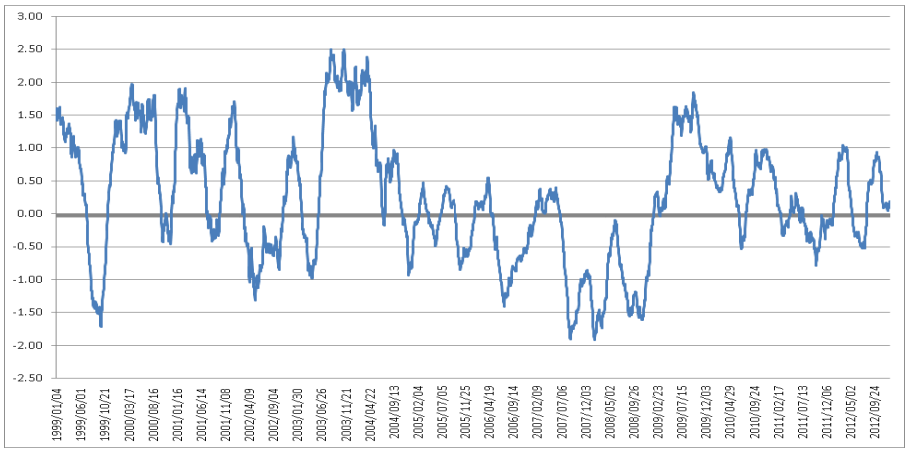

Figure 3.5 NPGR_60 from January 1999 to December 2012

외국인 투자자의 순매수금액과 마찬가지로 서포트 벡터 머신 알고리즘을 이용하여 구축되는 Lag ℓ 분류기의 입력변수들도 표준화 작업이 필요하기 때문에 각각의 입력변수들을 수정하였다. 즉, 일별 외국인 투자자의 순매수금액 (NPG)과 코스피200 지수 선물의 외국인 순매수금액 (NPI)은 코스피 시가총액으로 나뉜 비율을 사용하였으며, 환율 (FER)과 코스피지수 (SPI), 다우지수 (DJI)는 각각 차분한 값

들을 이용하였다. 그리고 수정된 원계열 값들에 대한 이동평균과 이동분산은 Table 3.5와 동일한 n-일을 사용하였으며, 외국인 투자자의 대량매도기간에서의 입력변수와 오라클 분류기를 통해 생성된 출력변수는 서포트 벡터 머신 알고리즘을 통한 Lag ℓ 분류기를 구축하였다. 본 실험에서도 Lag ℓ을 구축하기 위하여 사용된 학습 데이터셋을 제외한 데이터셋이 테스트 데이터셋으로 사용된다.

Table 3.7 Training and testing rates (%) of EWSFI trained by SVM

Year	Datasets	Y (SP, TP, CP)	Y (SP, CP)
1999	training	85.27	96.89
	testing	49.44	69.34
2002	training	73.46	81.63
	testing	48.65	61.13

구축된 EWSFI로 시장상황을 예측한 결과는 Table 3.7과 같다. 예측성과를 살펴보면, 첫 번째 실험한 결과와 유사하게 대량매도구간인 ESP99로 학습된 Lag ℓ 분류기가 학습 데이터셋과 테스트 데이터셋 모두 더 나은 예측성과를 보였다. 이러한 결과는 기계학습 알고리즘을 통한 예측방법이 학습 데이터셋의 적절한 선택이 중요한 부분임을 확인 할 수 있었다. 또한 시장상황을 3가지로 구분한 경우보다 2가지로 구분한 경우가 더 높은 예측성과를 보였다. 이와 같은 결과는 데이터의 표본이 커짐으로써 시장상황 3가지를 예측하는 경우에 불안정 (TP)을 예측하는 입력변수들의 특징들이 안정 (SP)과 위기 (CP) 구간들의 입력변수들의 특징들보다 명확하지 못한 것으로 판단된다. 즉, 노이즈로써 작용하여 앞에서 제시한 예측성과보다 좋지 못한 결과를 도출하였다. 이에 대한 결과는 시장상황을 2가지로 예측하는 경우를 통해 불안정 (TP)을 제외시킨 경우 평균적으로 65%의 시장상황을 올바르게 예측함을 보임으로써 외국인 투자자의 동향을 20일 전에 탐지하여 발생 가능한 위기를 조기에 예방 할 수 있는 조기경보시스템의 역할을 수행 가능할 것으로 판단된다.

4. 결론

본 연구는 기존의 Ahn 등 (2011)이 제안한 방법을 토대로 외국인 투자자의 국내주식시장의 순매수금액을 기준으로 구축된 오라클 분류기를 통해 시장상황을 판단하였으며, 서포트 벡터 머신 알고리즘 통하여 외국인 투자자의 순매수 금액에 따른 시장상황을 예측하는 지능형 조기경보시스템을 구축하였다. 본 연구에서는 Ahn 등 (2011)이 실험한 다른 알고리즘들보다 서포트 벡터 머신 알고리즘의 유용성을 확인할 수 있었으며, 2007, 2008년에 발생한 금융위기기간을 포함한 2012.12까지 구축된 EWSFI을 통하여 예측성과를 살펴 본 결과 외국인 투자자의 비정상적인 매도패턴을 사전에 미리 대비할 수 있는 지능형 조기경보시스템을 구축하였다. 또한, 본 연구에서 구축한 EWSFI는 국내주식시장뿐만 아니라 환율 및 원유시장 등 다양한 경제 분야에서 활용될 수 있는 가능성을 시사하고 있으며, 시장상황의 위기를 조기에 예측하여 예상되는 충격을 완화시킬 수 있다고 판단된다. 더 나아가 다양한 기계학습 알고리즘들을 혼합하여 EWSFI를 구축한다면, 장·단기 외국인 투자자의 비정상적 매도성향패턴을 파악한 지능형 조기경보시스템을 구축할 수 있기를 기대한다.

참고문헌

Ahn, J. J., Son, I. S., Oh, K. J., Kim, T. Y. and Song, G. M. (2011). Lag-ℓ forecasting and machine-learning algorithms. *Expert Systems*, **28**, 269-282.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167.

Choe, H., Kho, B. and Stulz, R. M. (1999). Do foreign investors destabilize stock markets? The Korean experience in 1997. *Journal of Financial Economics*, **54**, 227-264.

Edison, H. (2000). *Do Indicators of financial crises work? An evaluation of an early warning system*, Board of Governors of the FRS International Finance Discussion Paper, 675.

Eichengreen, B., Rose, A. K. and Wyplosz, C. (1995). Exchange market mayhem. *Economic Policy*, **21**, 251-312.

Eichengreen, B., Rose, A. K. and Wyplosz, C. (1996). *Contagious currency crises*, Working Paper 5681, National Bureau of Economic Research, London.

Frankel, J. and Rose, A. (1996). Currency crashes in emerging markets: An empirical treatment. *Journal of International Economics*, **41**, 351-366.

Ghysels, E. and Seon, J. (2005). The Asian financial crisis: The role of derivative securities trading and foreign investors in Korea. *Journal of International Money and Finance*, **24**, 607-630.

Goldstein, M., Kaminsky, G. and Reinhart, C. (2000). *Assessing financial vulnerability: An early warning system for emerging markets*, Institute for International Economics, Washington, D.C..

Kaminsky, G.L. and Reinhart, C. M. (1998). Financial crises in Asia and Latin America: Then and now. *American Economic Review*, **88**, 444-448.

Kaminsky, G.L. and Reinhart, C. M. (1999). The twin crises: The causes of banking and balance of payments problems. *American Economic Review*, **89**, 473-500.

Kim, K. and Kwon, S. (2003). *How has Korean economy changed during 5 years' financial crisis*? (in Korean), Samsung Economy Research Institute, Seoul, Korea.

Kim, W. and Wei, S. (1999) *Foreign portfolio investors before and during a crisis*, Working Paper 6968, National Bureau of Economic Research, Cambridge.

Krugman, P. (1979). A model of balance of payments crises. *Journal of Money, Credit and Banking*, **11**, 311-325.

Lapedes, A. and Farber, R. (1988). How neural networks work. In *Neural Information Processing Systems*, edited by D. Z. Anderson, American Institute of Physics, New York.

Lee, J. Y and Lee, J. H. (2010). Support vector machine and multifactor dimensionality reduction for detecting major gene interactions of continuous data. *Journal of the Korean Data & Information Science Society*, **21**, 1271-1280.

Obstfeld, M. (1986). Rational and self-fulfilling balance-of-payments crises. *American Economic Review*, **76**, 72-81.

Sachs, J. and Radelet, S. (1998). *The onset of the east asian financial*, Working Paper 8060, National Bureau of Economic Research, Cambridge.

Son, I. S., Oh, K. J., Kim, T. Y. and Kim, D. H. (2009). An early warning system for global institutional investors at emerging stock makrets based on machine learning forecasting. *Expert Systems with Applications*, **36**, 4951-4957.

Journal of the Korean Data & Information Science Society
2013, **24**(2), 223–233

http://dx.doi.org/10.7465/jkdi.2013.24.2.223
한국데이터정보과학회지

An intelligent early warning system for forecasting abnormal investment trends of foreign investors[†]

Kyong Joo Oh[1] · Young Min Kim[2]

[12]Department of Information and Industrial Engineering, Yonsei University

Received 20 November 2012, revised 11 December 2012, accepted 28 January 2013

Abstract

At local emerging stock markets such as Korea, Hong Kong, Singapore and Taiwan, foreign investors (FI) are recognized as important investment community due to the globalization and deregulation of financial markets. Therefore, it is required to monitor the behavior of FI against a sudden enormous selling stocks for the concerned local governments or private and institutional investors. The main aim of this study is to propose an early warning system (EWS) which purposes issuing a warning signal against the possible massive selling stocks of FI at the market. For this, we suggest machine learning algorithm which predicts the behavior of FI by forecasting future conditions. This study is empirically done for the Korean stock market.

Keywords: Abnormal investment trends, early warning system, foreign investors.

[†] This research was supported by 2010 LG Yonam Foundation funded by research grant for the overseas university professors.

[1] Corresponding author: Associate professor, Department of Information and Industrial Engineering, Yonsei University, Seoul 120-749, Korea. E-mail: johanoh@yonsei.ac.kr

[2] Ph.D. candidate, Department of Information and Industrial Engineering, Yonsei University, Seoul 120-749, Korea.