Journal of The Korea Society of Computer and Information
Vol. 21 No. 6, pp. 9-19, June 2016

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.6.009

Predicting Stock Liquidity by Using Ensemble Data Mining Methods

Eun Chan Bae *, Kun Chang Lee **

Abstract

In finance literature, stock liquidity showing how stocks can be cashed out in the market has received rich attentions from both academicians and practitioners. The reasons are plenty. First, it is known that stock liquidity affects significantly asset pricing. Second, macroeconomic announcements influence liquidity in the stock market. Therefore, stock liquidity itself affects investors' decision and managers' decision as well. Though there exist a great deal of literature about stock liquidity in finance literature, it is quite clear that there are no studies attempting to investigate the stock liquidity issue as one of decision making problems. In finance literature, most of stock liquidity studies had dealt with limited views such as how much it influences stock price, which variables are associated with describing the stock liquidity significantly, etc. However, this paper posits that stock liquidity issue may become a serious decision-making problem, and then be handled by using data mining techniques to estimate its future extent with statistical validity. In this sense, we collected financial data set from a number of manufacturing companies listed in KRX (Korea Exchange) during the period of 2010 to 2013. The reason why we selected dataset from 2010 was to avoid the after-shocks of financial crisis that occurred in 2008. We used Fn-GuidPro system to gather total 5,700 financial data set. Stock liquidity measure was computed by the procedures proposed by Amihud (2002) which is known to show best metrics for showing relationship with daily return. We applied five data mining techniques (or classifiers) such as Bayesian network, support vector machine (SVM), decision tree, neural network, and ensemble method. Bayesian networks include GBN (General Bayesian Network), NBN (Naive BN), TAN (Tree Augmented NBN). Decision tree uses CART and C4.5. Regression result was used as a benchmarking performance. Ensemble method uses two types-integration of two classifiers, and three classifiers. Ensemble method is based on voting for the sake of integrating classifiers. Among the single classifiers, CART showed best performance with 48.2%, compared with 37.18% by regression. Among the ensemble methods, the result from integrating TAN, CART, and SVM was best with 49.25%. Through the additional analysis in individual industries, those relatively stabilized industries like electronic appliances, wholesale & retailing, woods, leather-bags-shoes showed better performance over 50%.

▸ Keyword : Stock liquidity, Data-mining, Ensemble methods, decision making

• First Author: Eun Chan Bae, Corresponding Author: Kun Chang Lee
*Eun Chan Bae (eunchanbae@gmail.com), Department of Global Business Administration, Sungkyunkwan University, Seoul 03063, Republic of Korea
**Kun Chang Lee (kunchanglee@gmail.com), SKK Business School/SAIHST (Samsung Advanced Institute of Health Sciences & Technology), Sungkyunkwan University, Seoul 03063, Republic of Korea
• Received: 2016. 04. 11, Revised: 2016. 05. 18, Accepted: 2016. 06. 07.
• This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A3A2038108).

I. Introduction

세계은행에 따르면 대한민국의 GDP 대비 주식시장 규모는 96.5%로 일본의 61.8%, 중국의 44.9%(World Bank, 2012)에 비해 상당히 높은 비중을 차지하고 있다. 대한민국의 주식시장은 경제를 이끌어가는 성장 동력으로써의 역할을 하고 있으며 그만큼 대한민국 경제에서 주식거래는 중요하게 다루어져 왔다. 그런데 최근 한국경제는 유가하락 및 경쟁력 상실에 따른 불경기가 계속되고 있으며 최근 1년간 미국 S&P 500 지수가 13% 상승(Bloomberg, 2015)한 데 반해 한국 코스피 지수는 횡보합세를 유지(한국거래소, 2015)하고 있다. 이러한 상황에서 대기업들은 사내유보금을 축적하고 있으며[1], 개인 투자자들은 가계부채 급증, 고령화와 청년실업 등의 이유로 주식시장을 이탈하고 있다[2]. 또한 최근의 정치인 테마주 작전 사례는 자본의 주식시장 이탈을 더욱 가속화시키고 있다. 이러한 대한민국 주식시장의 규모 축소라는 문제 상황에서 최근 재무분야 연구자들은 주식시장 유동성 문제 및 기업 주식 유동성에 영향을 미치는 요소에 대한 연구를 진행하고 있다. 그러나 투자자나 회사 경영진의 의사결정에서 필요한 미래 주식 유동성 변화 방향에 대한 연구는 현재 진행되지 않고 있다. 따라서 본 연구에서는 실무적으로 투자자와 회사 경영진이 주식 유동성의 미래 움직임을 보다 정확하게 예측하도록 지원하여 주식시장 투자에 관한 의사결정을 지원하고자 한다.

주식 유동성이 높은 경우 투자자는 수익 실현 후 매도를 용이하게 할 수 있으나, 주식 유동성이 낮은 경우 추가 손해를 보고 손절해야 하는 경우가 생긴다[3]. 회사 경영진은 자금조달을 위한 의사결정을 내려야 하는데 자사의 주식 유동성이 높을수록 미청약에 대한 위험이 줄어들어 증자를 통한 자금조달 효과가 크다[4]. 그런데 기업 주식 유동성은 산업마다 다르고 기업마다 다르기 때문에 유동성 정도 및 변화 방향을 예측하는 것이 어렵다. 기존의 재무 관련 연구들에서는 주식 유동성이 주가에 미치는 영향에 대한 연구[5]와 특정 변수가 주식 유동성에 어떠한 영향을 미치는 가에 대한 연구[1]만이 진행되어 왔다. 그러나 여러 변수가 동시에 변할 때 주식 유동성은 어떻게 변화할지에 대한 연구는 단순 회귀분석 정도의 수준에서 머물고 있다.

따라서 본 논문에서는 주식 유동성에 관한 기존 연구의 한계인 단순 인과관계 확인 및 설명변수 파악 등의 수준을 데이터마이닝을 이용한 의사결정지원의 문제로 발전시키고자 한다. 또한 이와 같은 연구의도를 통하여 실무적으로 투자자와 회사 경영진이 주식 유동성의 미래 움직임을 보다 정확하게 예측하도록 지원하고자 한다. 그리고 이를 통하여 한국 주식시장이 활기를 되찾을 수 있는 계기를 마련하고자 한다. 이를 위하여 본 연구에서는 기존 연구에서 사용되어 온 회귀분석과 데이터마이닝 기법 간 주식 유동성 예측치에 대하여 비교하고(가설 1), 주식 유동성을 가장 잘 예측할 수 있는 데이터마이닝 방법에 대하여 연구하고자 한다(가설 2).

본 연구에서는 우리나라 한국거래소에 상장된 기업들의 주가 자료 및 회계 자료를 이용하여 기업의 주식 유동성 예측 모델을 구축하였다. 본 연구에서는 각 기업들의 공시된 회계 자료에 기반한 요소별 가치의 변동에 따라 주식 유동성에 어떠한 영향을 미치는지 분석하여 투자자들의 의사결정과 회사 경영진들의 의사결정에 도움을 줄 수 있는 주식 유동성 예측 모델을 구축하였다.

본 연구의 목적은 다음과 같다.

첫째, 데이터마이닝 기법을 이용하여 우리나라 기업의 주식 유동성 변화 예측모형을 제시한다.

둘째, 연구를 통해 얻은 데이터마이닝 모델을 통해 특정 산업의 기업 주식 유동성을 더욱 효과적으로 예측한다.

셋째, 데이터마이닝 기법을 이용한 주식 유동성 예측모델을 개발함으로서 투자자와 회사 경영인의 의사결정에 도움을 주는 방법으로서의 지평을 확장한다.

본 연구의 구성은 다음과 같다. 제 Ⅱ장에서는 기업의 주식 유동성에 관한 기존 선행 연구를 간략하게 소개한 후, 본 연구에서 검증하기 위한 연구목표, 가설 및 데이터마이닝 방법에 대하여 서술하였다. 제 Ⅲ장에서는 기업의 주식 유동성 예측을 위해 사용된 변수들과 데이터마이닝 방법에 대하여 서술하였다. 제 Ⅳ장에서는 제 Ⅲ장에서 언급된 데이터마이닝 방법을 가지고 기업의 주식 유동성을 예측해보고, 앙상블 방법을 통하여 더욱 예측력이 우수한 모델을 개발하였다. 또한 구축된 모델을 이용하여 특정 산업의 주식 유동성 예측 정도를 확인하였다. 끝으로 Ⅴ장에서는 본 연구의 시사점과 한계점 및 연구방향을 제시하였다.

II. Preliminaries

1. Related works

1.1 기업 주식 유동성에 관한 선행 연구

기업 주식 유동성에 관한 선행 연구는 글로벌 금융위기 이후 투자자와 회사 경영진의 의사결정에 영향을 미치는 요소로서 연구가 진행되어왔다. 최근에는 기업 주식 유동성 자체에 어떠한 요소가 영향을 미치는 지에 대한 연구가 재무 및 의사결정 분야에서 진행되고 있다.

1.1.1 독립변수로서의 주식 유동성 관련 연구

투자자의 의사결정에 영향을 미치는 요소로서 주식 유동성 관련 연구는 주식 유동성이 주식 수익률에 영향을 미치는가에 대한 연구를[6] 시작으로 주식 유동성 프리미엄에 대한 연구가 진행되어 왔으며 한국에서는 기업의 주식 유동성이 주가에 미치는 영향[5]에 대한 연구가 진행되어왔다. 또한 주식 유동성이 높을수록 주식 브로커의 대량주문 가능성이 많아진다[7]는 연구가 진행되어왔다. 회사 경영진의 의사결정에 영향을 미칠 수 있는 요소로서 주식 유동성 관련 연구는 기업의 주식 유동성이 높을수록 재무구조 레버리지 비율이 낮아서 주식시장에서 자금조달을 하는 경우가 많다는 기업의 주식 유동성과 기업 재무 구조의 관계[4]에 대한 연구와 기업의 주식 유동성이 높을수록 기업의 자금조달비용이 낮다[8]는 연구가 진행되어왔다.

1.1.2. 종속변수로서의 주식 유동성 관련 연구

기업의 주식 유동성에 어떠한 요소가 주식 유동성에 영향을 미치는 지에 대한 연구는 글로벌 금융위기 이후 연구가 활발해지기 시작하였다. 우선 기업이 보유하고 있는 유동성 자산을 비유동성 자산으로 바꿀 때 주식 유동성이 감소한다[9]는 연구가 진행되었으며, 이 연구가 한국의 주식시장에서도 적용하는지에 대한 연구가 진행되었다 [1]. 또한 특정 기업의 주식을 대량 보유하고 있는 내부자 및 외부자가 증가할수록 그 기업의 주식 유동성이 감소한다는 연구가 진행되었다 [10].

Table 1. Related works and its validating model

Authors, Year	Main contents	Validating model
Previous research that denoted stock liquidity as an independent variable		
H.J.Ko, 2009[5]	Independent variable: Stock liquidity variance Dependent variable: Additional profit Result: Stock liquidity variance is inversely proportional to additional profit	Regression analysis
Lipson et al., 2009[4]	Independent variable: Stock liquidity Dependent Variable: Corporate financial structure Result: Stock liquidity is inversely proportional to corporate financial leverage ratio	Panal-data analysis
Kryzanowski & Lazrak, 2010[8]	Independent variable: Stock liquidity Dependent Variable: Cost of capital Result: Stock liquidity is inversely proportional to cost of capital	Regression analysis
Turnbull et al., 2010[7]	Independent variable: Stock liquidity Dependent Variable: Stock broker`s bulk order Result: Stock liquidity is proportional to stock broker`s bulk order	Regression analysis
Previous research that denoted stock liquidity as a dependent variable		
Gopalan et al., 2012[9]	Independent variable: Corporate liquidity asset Dependent Variable: Stock liquidity Result: Corporate liquidity asset is proportional to stock liquidity	Regression analysis
K.S.Cho, 2013 [10]	Independent variable: The number of people who holds bulk of stocks Dependent Variable: Stock liquidity Result: The number of people who holds bulk of stocks is inversely proportional to stock liquidity	Regression analysis
H.C.Lee, 2014[1]	Independent variable: Corporate liquidity asset in korean company Dependent Variable: Stock liquidity Result: Corporate liquidity asset in korean company is proportional to stock liquidity (Verify Gopalan et al.`s idea in korean market)	Regression analysis

선행 연구들의 주식 유동성에 관한 가설 검증에는 변수간의 관계를 보기 위한 회귀분석, 변수들의 특징을 보기 위한 패널 분석 등이 사용되었다. 그러나 이 모델들은 변수간의 상관관계만을 보기 위한 방법으로, 주식 유동성이 어떻게 변화할 것인지를 예측하는데 있어서는 한계가 있다. 또한 회귀분석의 경우 선형성이라는 강력한 제한 때문에 실제 값을 정확하게 예측하는 데에는 한계가 있다. 따라서 본 연구에서는 회귀분석 및 패널분석 방법 이외에 데이터마이닝 기법인 베이지안 네트워크(Bayesian Network), 의사결정트리(Decision Tree), 서포트 벡터 머신(Support Vector Machine), 인공신경망(Artificial Neural Network)을 이용하였고 다양한 데이터마이닝 기법을 조합할 수 있는 앙상블 방법을 사용하였다.

1.2 데이터마이닝 방법 관련 연구

본 연구에서 사용한 데이터마이닝 기법은 다음과 같다.

1.2.1 베이지안 네트워크(Bayesian network)

베이지안 네트워크는 비순환 방향성 그래프 구조를 가진 확률모델로서 조건부 확률 테이블을 이용하여 변수들 사이의 인과관계를 표현한다[11]. 이 때에 변수는 노드(Node)로 표현되며 변수들 간의 인과관계는 노드 간 연결된 아크(Arc)를 통한 확률적 인과관계로 표현한다. 이 때, 주어진 의사결정 문제의 영역지식을 확률적으로 표현하여 구성 변수들간에 존재하는 확률적 의존관계의 방향을 아크(arc)로 나타냄으로써 각 변수들의 조건부 확률을 계산하고 인과관계를 표현한다[12]. 베이지안 네트워크는 경영학 연구에서 의사결정을 지원하는 방법으로 유용하게 사용되어 왔으며 특히 베이지안 네트워크를 이용한 주가 예측[13]과 같은 재무분야 의사결정에서도 유용한 것으로 확인되었다.

베이지안 네트워크는 가장 단순한 형태인 NBN(Naive Bayesian Network),이로부터 확장된 형태인 TAN(Tree Augmented Naive Bayesian Network) 그리고 가장 일반적인 형태인 GBN(General Bayesian Network)이 있다.

NBN(Naive Bayesian Network)는 네트워크 내의 모든 노드 혹은 특성들이 주어진 클래스 내에서 서로 독립이라는 가정을 한다. 이러한 독립 가정 때문에 속성의 수가 많을 때 각 속성의 모수들을 분리하여 학습을 간단하게 한다. 다만, 변수들이 서로 독립이라는 가정은 대부분의 현실 세계의 문제를 제대로 반영하지 못한다는 단점이 있다[14].

TAN(Tree Augmented Naive Bayesian Network)은 NBN과는 다르게 노드 혹은 특성들 간 상호의존도가 존재한다고 가정하고 이러한 의존도를 베이지안 네트워크 형태로 표현 가능하도록 나이브 베이지안 네트워크(NBN)방법을 확장한 것이다. TAN은 변수들 간 의존성을 추가함으로서 통해 NBN의 학습능력을 향상시켜 분류성능을 높인다[15].

GBN(General Bayesian Network)는 타겟 노드에 해당하는 목표 변수를 일반 노드와 동일하게 다루는 베이지안 네트워크이다. GBN에서는 타겟 노드도 부모 노드들을 가질 수 있기 때

문에 주어진 목표와 관련된 여러 변수 간에 존재하는 확률적 인과관계를 자연스럽게 표시할 수 있다. 이러한 방법은 실제 현실 세계를 잘 반영한다는 이유로 다양한 의사결정 연구에서 사용되어왔다. 부동산 가격 예측과 같은 의사결정에서 GBN은 유용한 것으로 확인된 것뿐만 아니라[16] 주가 예측 분야에서도 GBN은 유용한 데이터마이닝 기법으로 사용되었다[13].

이처럼 모든 변수를 동일하게 다루는 GBN을 형성하기 위해 여러 방법의 알고리즘이 사용되고 있다. 그 중 GBN-K2는 주어진 일련의 속성들로 시작하여 기존의 노드로부터 다른 노드로 선을 뻗어 가는데, 매 단계마다 검증력을 최대화하고자 한다. 더 이상의 개선이 없으면 다음 노드로 진행되며 과적합 문제(overfitting)을 막고자 각각의 부모 노드의 수는 사전에 정해진 최대 수만큼 제한된다. 이 과정에 의하여 비-순환 구조가 형성된다[17]. 또 다른 알고리즘은 GBN-HC (Hill Climb)이다. 비어있거나 임의의 네트워크에서 시작하여 매번 가능한 모든 경로(추가, 삭제, 반전)들을 고려하여 성능을 가장 향상시키는 방법을 선택하게 된다[18].

1.2.2 의사결정트리(Decision tree)

의사결정트리(Decision tree)는 불분명한 변수의 성격이나 변수간의 관계를 지닌 대용량의 자료를 처리할 때 사용되는 분류방법이다[19]. 의사결정나무는 특히 정보를 쉽게 이해할 수 있다는 장점이 있는데 루트 노드에 존재하는 변수에서 시작하여 나머지 하위 노드들간에 'If-Then'의 관계가 설정되어있기 때문이다. 따라서 의사결정트리는 데이터베이스로부터 데이터를 분석, 패턴을 분석하는 다양한 분야에서 효과적인 도구들로 인식되고 있다[20]. 본 논문에서는 의사결정트리 방법 중 의사결정분야에서 널리 쓰이고 있는 CART 알고리즘과 C4.5 알고리즘을 사용하여 주식 유동성을 예측하였다.

CART 알고리즘은 분할규칙에 근거하여 의사결정트리를 생성하는데, 이 분할규칙의 기본 아이디어는 모든 가능한 분할 중에서 가장 순수한 결과의 자식 노드가 얻어지도록 하나의 분할을 선택하는 특징이 있다[21].

C4.5 알고리즘은 CART 알고리즘과 유사하나 가지의 수를 다양화 할 수 있다는 점에서 CART 알고리즘의 이진분리와는 차이가 있다. 또한 가지치기 방법도 CART와는 다르게 훈련 데이터와 멀리 떨어져있는 데이터에 대해서는 언급하지 않고 가지치기를 할 때에 훈련 데이터와 같은 데이터를 적용한다[19].

1.2.3 서포트 벡터 머신(SVM)

서포트 벡터 머신은 클래스 분류와 특징 선택에 있어서 유용한 분류기이다[22]. SVM은 데이터를 분리하는 초평면 중에서 데이터들과 거리가 가장 먼 초평면을 선택하여 분리하는 방법이다[23]. 초기에는 선형 결합에 가중치를 부여하는 방식을 통해 이진분류 문제를 위해 설계되었다. 만약 x 개의 클래스가 있다면, 각각의 SVM은 모든 클래스에 대해 학습하며, X(X-1)/2 개의 클래스는 멀티 SVM 분류기를 구성하기 위해 사용된다. 두 클래스를 분류하는 초평면은 다른 클래스 간 거리가 가장 가까운 점이 다른 초평면들과 비교하였을 때는 그 거리가 가장 긴 초평면으로 사용한다[24]. 최근에는 SVM을 다중 클래스 문제에 적용하기 위한 방법으로 다중 이진 분류기의 출력 결과를 이용하는 출력 코딩 방법이 주로 사용되고 있다. 본 논문에서는 다중 클래스가 사용되었기에 다중 이진 분류기의 출력 결과를 이용하는 방식을 사용하였다.

1.2.4 인공신경망(Artificial neural network)

인공신경망 방법은 사람 뇌의 기본 신경 단위인 뉴런(Neuron)을 모방한 방법이다[25]. 일반적인 인공신경망은 입력층(Input Layer), 은닉층(Hidden Layer), 그리고 출력층(Output Layer)로 이루어져있다. 각각의 층은 각 노드들의 집합으로 구성되며, 노드 간에는 연결 가중치가 부여되어 상호 연결되어있다. 각각의 노드는 전 단계의 출력값을 입력값으로 받아 내부에서 생성된 함수에 의해 출력값을 생성한다. 인공신경망은 학습 데이터를 통해 노드 간의 가중치를 결정하고 이를 가지고 실제 데이터 예측을 하게 된다. 인공신경망은 서비스 수요예측[26], 신용도 예측[27] 등 다양한 의사결정분야에서 사용되고 있다.

1.2.5 앙상블 학습 방법

앙상블 학습 방법은 어떠한 분류기도 다른 분류기에 비해 언제나 좋은 성능을 보인다고 말할 수 없다는 점에서 착안한다[28]. 그래서 여러 개의 분류기를 합쳐서 사용하는 것이 최종적인 분류에 더 많은 훈련 데이터에 대하여, 더 나은 예측력을 보인다는 점에서 앙상블 학습 방법은 사용되고 있다[29]. 앙상블 학습 방법에 대하여 여러 개의 분류기를 합치는 것은 단일 분류기를 사용하였을 때에 나타날 수 있는 오류를 줄인다는 점에서 더 좋은 학습 방법으로 알려져 있다[30]. 앙상블 학습 방법은 개별 분류기를 가지고 데이터의 학습 자료 및 실험 자료를 재구성하여 다양한 데이터의 특징을 반영할 수 있는 Bagging, Boosting방법과 각각의 훈련용 데이터 예측결과와 각각의 분류기에서 나온 예측력을 가중치로 환산하여 사용하는 Voting방법이 있다[31]. 앙상블 방법들은 그 목적에 맞게 다양하게 사용되고 있다. 의사결정 연구에서는 부도예측을 위한 통합알고리즘[32] 개발 등 데이터마이닝 모델의 예측력을 높이기 위한 모델로서 Voting 방법이 많이 쓰이고 있다. 따라서 본 연구에서는 앙상블 모델 방법 중 Voting방법을 사용하기로 한다.

III. The Proposed Scheme

기업의 주식 유동성을 효과적으로 예측하고, 이를 투자자와 회사 경영진의 의사결정에 실무적으로 활용하기 위하여 다양한

데이터마이닝 기법을 이용하였다. 본 연구에서는 기업의 주식 유동성 예측을 위한 모델을 구축 및 실제 의사결정 지원을 위하여 다음과 같은 연구 목표를 설정하였다.

Study 1

Develop a model to predict stock liquidity

1.1 가설 1. 회귀분석보다 데이터마이닝 기법이 기업의 주식 유동성을 더 잘 예측할 수 있을 것이다.

선행 연구에서 지속적으로 사용되어온 회귀분석은 변수의 유의미성을 판단하는 데 있어서는 좋은 분석도구이다. 그러나 모든 데이터를 직선으로 표현하는 선형성의 한계 때문에 미래의 값을 정확하게 예측하는 데에는 한계가 있다. 연구 1에서는 회귀분석 방법의 한계를 극복할 수 있는 다양한 데이터마이닝 기법에 대해 한국거래소의 자료를 가지고 실증 분석하였다. 그리고 이를 통해 기업 주식 유동성 예측 모델을 개발하였다. 연구에 따르는 가설 1은 다양한 데이터마이닝 기법을 가지고 회귀분석보다 주식 유동성 예측력이 좋은 모델을 확인한다.

1.2 가설 2. 데이터마이닝 기법 중 앙상블 방법이 주식 유동성 예측력을 더욱 증가시킬 수 있을 것이다.

주가는 다양한 원인에 의하여 그 변화의 폭이 심하다. 그리고 주식 가격에 따라 변화하는 주식 유동성 또한 변화 과정을 예측하기 힘들다. 이러한 상황 속에서 단일 분류 방법을 이용한 유동성 예측은 현실 세계의 다양한 상황을 고려하지 못한다는 점에서 한계가 있다. 이 때 다양한 분류 모델을 합하여 예측을 실시하는 앙상블 방법을 사용한다면 복잡한 현실을 반영하여 주식 유동성 예측을 더 잘 할 수 있다. 또한 기존 연구에서도 앙상블을 사용한 방법이 단일 분류 방식보다 더 좋은 예측력을 보임[30]을 확인하고 있다. 따라서 가설 2에서는 추가적으로 데이터마이닝 방식 중 예측력이 좋은 모델을 중심으로 앙상블 기법을 적용하여 더 좋은 예측력을 가질 수 있는지 확인한다. 그리고 이를 통해 최종적으로 기업의 주식 유동성을 예측하는 모델을 발견한다.

Study 2

Stock liquidity prediction analysis between industry

본 연구에서는 각 회사의 재무 특성을 반영하는 변수들로 기업의 주식 유동성을 측정하고자 한다. 그런데 이러한 재무 특성은 각 산업별로 차이가 있다. 예를 들어 건설업의 대규모 공사에 들어가는 비용과 서비스업에서 광고에 들어가는 비용은 분명 차이가 있고, 이에 따른 재무적 특성 자체도 다르다. 따라서 연구 2에서는 연구 1을 통하여 구축된 모델을 가지고 산업 군을 나누어 모델을 실제로 적용해보았다. 그리고 특정 산업에서 모델의 예측력의 정도를 확인하였다. 또한 이를 통해 투자자와 회사 경영자의 의사결정에 도움이 될 수 있는 방안에 대하여 논의해 보았다.

Ⅳ. Experiment Method

1. Data and dependent variables

본 연구에서는 한국거래소에 상장된 기업 중 2010년부터 2013년까지 기간의 기업을 표본으로 하였다. 2010년 이후로 기간을 정한 것은 글로벌 금융위기로 인한 특수한 상황을 배제하려고 했기 때문이다.

Fn-GuidePro로부터 결산기가 12월말인 기업에 한정하여 재무자료, 주가자료 등 자료가 수집 가능한 기업들을 대상으로 하였다. 그래서 총 1,518사, 5,700개의 기업-연도 표본을 선정하였다.

종속변수인 기업 주식 유동성 측정치는 Amihud측정치를 이용하였다. Amihud측정치는 주식 유동성을 측정하기 위한 측정치[6]로 고안되었으며 한국의 주식시장에서 주식 유동성을 잘 반영하는 것[33]으로 확인되었다. 또한 Amihud 측정치는 주식 유동성 측정치가 만족해야 할 여러 가설들을 만족하고 있다는 사실이 확인되었다[1]. 따라서 본 연구에서는 주식 유동성 측정을 위해 Amihud측정치를 사용하기로 하였다.

Amihud 측정치는 Amihud 측정치는 일별 수익률의 절대값을 거래대금으로 나누어 구한다.

$$Amihud = \frac{1}{T}\sum_{t=1}^{T}\frac{|\text{일별수익률}_t|}{\text{거래대금}_t}$$
$$where\ t = \text{거래일}$$

분석에 사용한 Amihud 측정치는 일별 측정치를 거래기간으로 평균하여 사용하였고, 분석의 편의를 위하여 측정치의 10^9를 곱하여 분석을 진행하였다. Amihud 측정치는 수익률에 거래금액 기준의 거래량으로 나눈 측정치로서, Amihud 측정치값이 낮을수록 주식의 유동성이 커지는 특징이 있다. 거래대금의 경우는 매일의 거래량에 당일 종가를 곱하여 구하였다.

2. Independent variable

독립변수로는 주식 유동성에 영향을 미치는 요소와 관련된 기존 연구[1]에서 확인한 주식 유동성에 큰 영향을 미치는 요소를 중심으로 자산유동성(Weighted asset liquidity, 이하 WAL), 시가총액(Market value of equity, 이하 MCAP), 성장기회의 크기가 주식 유동성에 미치는 영향을 통제하는지 확인하기 위한 자본지출(Capital expenditure, 이하 CAPX), 시장가치-장부가치 비율(Prie book ratio, 이하 PBR), 기업의 성과와 주식 유동성 간의 영향을 보기 위한 자산대비수익률(Return on asset, 이하 ROA), 주식 보유기간 중 초과수익(Buy-and-hold annual abnormal stock return, 이하 BHAR), 기업의 투명한 공시에 주식 유동성이 미치는 영향을 보는 재량적 발생액(Discretionary Accruals, 이하 DA), 수익률의 변동성(Volatility, 이하 VOL)을 독립변수로 사용하였다.

자산유동성(WAL)은 유동성 자산에 대해 가중평균을 한 자산 유동성 추정 방식을 사용하였다[9]. 시가총액(MCAP)은 시

가총액에 자연 로그 값을 취하여 구하였다. 자본적지출(CAPX)은 토지의 가격을 베외한 고정자산 증가분에 감가상각비를 더하였다. 시장가치-장부가치 비율(PBR)은 시가총액을 자기자본의 장부가치로 나누어 구하였다. 자산대비수익률(ROA)은 순수익을 총자산으로 나누어 구하였다. 주식 보유기간 중 초과수익(BHAR)은 거래기간 중 각 종목의 수익률에 KOSPI지수 수익률을 차감하여 구하였다. 수익률의 변동성(VOL)은 거래기간 중 수익률의 표준편차에 자연로그 값을 취하여 구하였다. 재량적 발생액(DA)은 기업의 공시 투명성 정도를 통제하기 위해 사용하는데, 수정존스모형[34]을 이용하여 총 발생액(accrual)에서 비 재량적 발생액 추정치를 하여 추정하였다. 재량적 발생액 추정을 위하여 FnGuide에서 찾을 수 있는 한국표준산업분류상 중분류를 기준으로 표본수가 산업 내 10개 이하인 경우는 제외하고 연도별·산업별 회귀모형을 통해 회귀계수를 추정하였다. 단, 조선 산업은 우리나라 주요 산업 분석을 위하여 포함시켰다. 이렇게 추정된 회귀계수를 이용하여 총 발생액에서 비 재량적 발생액의 크기를 감하여 재량적 발생액의 크기를 추정하였다.

Table 2. Descriptive Statistics

Variable	Average	Standard deviation	Lower 10%	Median	Upper 10%
Amihud	1.7695	12.9337	0.005	0.069	2.045
WAL	0.322	0.150	0.155	0.308	0.499
MCAP	23.89	4.362	16.04	24.82	27.04
CAPX	$1.089*10^8$	$9.454*10^8$	211805	565881	$7.89*10^8$
PBR	1.3259	1.8490	0.3715	0.9378	2.5879
ROA	0.000674	0.2197	-0.105	0.0276	0.10401
BHAR	-0.06	0.439	-0.55	-0.03	0.455
DA	$-3.41*10^7$	$5.00*10^8$	$-4.5*10^7$	$-1.7*10^6$	$1.24*10^7$
VOL	-3.05	1.267	-3.93	-3.40	0.282

3. Hypothesis validity model

연구 가설을 검증하기 위하여 종속변수는 주식 유동성을 측정하는 Amihud 측정치를 사용하였고 독립변수는 유동성 자산(WAL), 시가총액(MCAP), 자본지출(CAPX), 시장가치-장부가치 비율(PBR), 자산대비수익률(ROA), 보유기간 추가 수익률(BHAR), 재량적 발생액(DA), 변동성(VOL)을 사용하였다. 본 연구를 위하여 WEKA 3.6.12 버전을 사용하였다.

연구 가설을 검증하기 위하여 독립변수들 내의 크기에 따라 각각 동일한 비중을 갖는 다섯 단계로 값을 분리하여 시행하였다. 베이지안 네트워크의 경우 독립변수들을 범주형으로 분리할 때에는 가우지안 분포를 사용하고[35] SVM은 주어진 벡터 값을 가지고 초평면을 분리하는 방식이기 때문에[23] 범주형으로 분리 할 필요가 없다. 그러나 앙상블 방법의 경우 두 개 이상의 데이터마이닝 방법을 조합하는 것으로서, 각각의 데이터마이닝 방법에 맞는 독립변수 분리를 동시에 실시할 수 없다. 따라서 본 연구에서는 변수들 내의 크기에 따라 각각 동일한 비중을 갖는 다섯 단계로 값을 분리하여 실험을 시행하였다.

Amihud 측정치의 경우 수익률에 거래금액 기준의 거래량으로 나눈 측정치로서, Amihud 측정치가 낮을수록 주식의 유동성이 커지는 특징이 있다. 이를 반영하여 Amihud 측정치의 범주를 다섯 가지로 나누었다. 측정치의 클래스 구분 값은 다음과 같다.

Table 3. Categories of Amihud
(Categories of Stock liquidity measurement depending on its value)

Amihud	Categories of Amihud
$(-\infty, 0.0136949]$	Very high
$(0.0136949, 0.0439251]$	High
$(0.0439251, 0.1148599]$	Middle
$(0.1148599, 0.5691508]$	Low
$(0.5691508, \infty)$	Very low

Ⅴ. Experiment

Study 1

Develop a model to predict stock liquidity

1.1 가설 1. 회귀분석보다 데이터마이닝 기법이 기업의 주식 유동성을 더 잘 예측할 수 있을 것이다.

Table 4. Results of stock liquidity prediction accuracy using data-mining methods and logistic regression

Data-mining methods using single classifier	Accuracy(%)
Regression(Logistic)	46.99
GBN-K2	47.42
GBN-HC	47.8
NBN	47.41
TAN	48.12(*)
C4.5	46.97
CART	48.42(**)
SVM	47.84
Artificial neural network	46.4

회귀분석 방법 중 범위형 자료를 분류하는 로지스틱 회귀분석을 기준으로 하여 베이지안 네트워크 계열의 GBN-K2, GBN-HC, NBN, TAN, 의사결정트리 계열의 C4.5과 CART, SVM 방법 및 인공신경망을 사용하였다. 이 때 분류모형 및 파라미터 사전설정은 Weka 3.6.12버전에 설정되어 있는 디폴트 값으로 처리하였다. 비록 추정 파라미터의 설정을 변함에 따라 예측력에 큰 차이가 생길 수 있지만 본 연구의 초점과는 다소 동떨어지기

에 추후 연구 문제로 남겨둔다. 예측력에 대한 모델 간 비교는 Weka 프로그램에 있는 T-검정 방법을 이용하였다.

이들 모든 데이터마이닝 기법들은 분석 결과 회귀분석보다 더 높은 예측력를 나타내고 있다. 또한 CART방법(**p<.05), TAN 방법(*p<.10)은 유의미한 예측력 차이를 보이고 있다. 이상의 결과로 알 수 있는 바로는 회귀분석은 비록 변수들의 상관관계를 확인하는 데에는 좋은 역할을 하지만, 주식 유동성을 예측하는 측면에서는 성능이 떨어짐을 확인할 수 있었다.

다만, table 3의 결과에서는 모든 예측력이 50%를 넘지 못하고 있는데, 이러한 결과는 실무에서 사용하는 데 제한이 있다. 이에 다음 가설에서는 앙상블방법을 이용하여 예측력을 좀 더 높이고자 한다.

1.2. 가설 2. 데이터마이닝 기법 중 앙상블 방법이 주식 유동성 예측력을 더욱 증가시킬 수 있을 것이다.

Table 5. Results of stock liquidity prediction accuracy using 2-ensemble data-mining methods

Ensemble methods using two classifiers		Accuracy(%)
Criteria : CART (Accuracy : 48.42%)		
GBN-K2	GBN-HC	48
GBN-K2	NBN	48.05
GBN-K2	TAN	48.45
GBN-K2	C4.5	47.12
GBN-K2	CART	48.44
GBN-K2	SVM	48.29
GBN-K2	Artificial neural network	47.76
GBN-H2	NBN	48.58
GBN-H2	TAN	48.5
GBN-H2	C4.5	47.52
GBN-H2	CART	48.6
GBN-H2	SVM	48.96
GBN-H2	Artificial neural network	47.95
NBN	TAN	48.61
NBN	C4.5	47.56
NBN	CART	48.71
NBN	SVM	48.13
NBN	Artificial neural network	48
TAN	C4.5	48.1
TAN	CART	48.99
TAN	SVM	48.94
TAN	Artificial neural network	47.84
C4.5	CART	47.52
C4.6	SVM	47.54
C4.7	Artificial neural network	47.2
CART	SVM	48.55
CART	Artificial neural network	47.94
SVM	Artificial neural network	47.41

2개의 분류기를 사용한 앙상블 방법에서는 'TAN+CART' 방법이 가장 우수한 예측력을 보였으나 단일 분류기에서 가장 좋은 예측력을 보인 'CART' 방법과 비교하여 유의미한 차이를 보이지는 않았다(p>0.10).

Table 6. Results of stock liquidity prediction accuracy using 3-ensemble data-mining methods

Ensemble methods using three classifiers			Accuracy(%)
Criteria : CART (Accuracy : 48.42%)			
GBN-K2	GBN-HC	NBN	48.5
GBN-K2	GBN-HC	TAN	48.41
GBN-K2	GBN-HC	C4.5	47.82
GBN-K2	GBN-HC	CART	48.63
GBN-K2	GBN-HC	SVM	48.72
GBN-K2	GBN-HC	Artificial neural network	48.44
GBN-K2	NBN	TAN	48.48
GBN-K2	NBN	C4.5	47.83
GBN-K2	NBN	CART	48.72
GBN-K2	NBN	SVM	48.77
GBN-K2	NBN	Artificial neural network	48.71
GBN-K2	TAN	C4.5	48.04
GBN-K2	TAN	CART	48.83
GBN-K2	TAN	SVM	48.84
GBN-K2	TAN	Artificial neural network	48.63
GBN-K2	C4.5	CART	47.85
GBN-K2	C4.5	SVM	47.59
GBN-K2	C4.5	Artificial neural network	48.07
GBN-K2	CART	SVM	48.75
GBN-K2	CART	Artificial neural network	48.72
GBN-K2	SVM	Artificial neural network	48.5
GBN-HC	NBN	TAN	49.09
GBN-HC	NBN	C4.5	48.28
GBN-HC	NBN	CART	49
GBN-HC	NBN	SVM	49.14
GBN-HC	NBN	Artificial neural network	49.01
GBN-HC	TAN	C4.5	48.3
GBN-HC	TAN	CART	49.04
GBN-HC	TAN	SVM	49.08
GBN-HC	TAN	Artificial neural network	48.62
GBN-HC	C4.5	CART	48.19
GBN-HC	C4.5	SVM	48.07
GBN-HC	C4.5	Artificial neural network	48.01
GBN-HC	CART	SVM	49.07
GBN-HC	CART	Artificial neural network	48.73
GBN-HC	SVM	Artificial neural network	48.47
NBN	TAN	C4.5	48.48
NBN	TAN	CART	49.07
NBN	TAN	SVM	49.02

NBN	TAN	Artificial neural network	48.81
NBN	C4.5	CART	48.26
NBN	C4.5	SVM	47.89
NBN	C4.5	Artificial neural network	47.21
NBN	CART	SVM	48.65
NBN	CART	Artificial neural network	48.82
NBN	SVM	Artificial neural network	48.57
TAN	C4.5	CART	48.41
TAN	C4.5	SVM	48.48
TAN	C4.5	Artificial neural network	48.3
TAN	CART	SVM	49.25(*)
TAN	CART	Artificial neural network	48.82
TAN	SVM	Artificial neural network	48.66
C4.5	CART	SVM	47.85
C4.5	CART	Artificial neural network	48.25
C4.5	SVM	Artificial neural network	47.84

3개의 분류기를 사용한 앙상블 방법에서는 'TAN+CART+SVM' 방법이 가장 좋은 예측력을 나타내었다. 또한 단일 분류기를 사용하였을 때 가장 좋은 결과를 나타낸 'CART'의 예측력과 유의미한 차이($p<0.10$)을 보였다. 앙상블 기법을 통하여 단일 데이터마이닝 모델보다 더 좋은 예측력을 보이는 모델을 얻을 수 있었다. 다만 결과에 대한 인과관계를 시각적으로 확인할 수 없었다. 이는 베이지안 네트워크나 의사결정트리 방법의 경우 간단한 그래프와 노드를 가지고 인과관계를 포함할 수 있으나 앙상블 방법은 여러 방법을 조합하는 것이기 때문에 그 결과를 확인할 수 없기 때문이다.

'TAN+CART+SVM' 방법의 Sensitivity, Specificity, F-1 value 및 ROC 커브는 아래와 같다.

Table 7. Detailed results of 'TAN+CART+SVM' including sensitivity, specificity, f-1 value and roc area

Class	Sensitivity	Specificity	F-Measure	ROC Area
VERY LOW	0.697	0.766	0.73	0.93
LOW	0.399	0.401	0.4	0.756
MIDDLE	0.321	0.284	0.302	0.711
HIGH	0.372	0.379	0.376	0.748
VERY HIGH	0.621	0.615	0.618	0.887

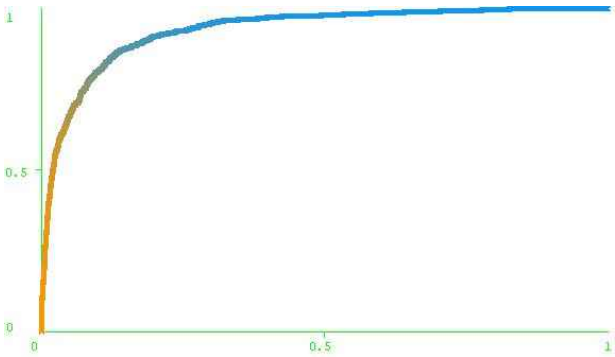


Fig. 1. Roc curve when class is 'very low' (Area under ROC= 0.9305)

Study 2

Stock liquidity prediction analysis between industry

연구 1을 통해 확인한 'TAN+CART+SVM' 모델을 가지고 산업군을 나누어 주식 유동성 예측을 실시하였다. 그리고 특정 산업에서의 예측력을 확인하였다. 본 연구에서 사용한 총 1,518개의 회사 중 각각의 산업에 해당하는 데이터의 변수는 다음과 같으며 연구 2를 위하여 2013년에 해당하는 산업별 데이터를 가지고 주식 유동성을 예측하였다.

Table 8. Descriptive statistics in each industry

Descriptive statistics		Number of sample	Average of stock liquidity (Amihud)
Manufacturing	Food	98	0.504
	Chemicals and chemical products (Except pharmaceuticals)	44	0.601
	Rubber and plastic	71	1.012
	Leather, bag and shoes	85	0.8226
	Wood and wood products	72	0.8504
	Petroleum	142	1.568
	Pulp, paper and paper-baesd manufacturing	229	0.351
	Printing and recorded media replication	31	0.9299
	Tobacco	82	0.1901
	Furniture	34	1.2145
	Clothing, Clothing accessory and fur clothing	32	0.8072
	Textile(Except clothing)	40	0.1546
	Automobile and trailer	19	0.5886
	Electrical and electronics	46	15.73
	Semiconductor	67	10.71
	Shipbuilding	7	0.5146
Service		58	1.51
Wholesale and retail trade		55	0.248
Finance and insurance		40	0.822
Transportation		147	0.594
Mining		92	0.655
Construction		27	9.84

Table 9. Results of stock liquidity prediction accuracy using 'TAN+CART+SVM' in each industry

Stock liquidity prediction accuracy in 2013 (TAN + CART + SVM)	Accuracy (%)
Electrical and electronics	60.86
Leather, bag and shoes	52.67
Wood and wood products	52.37
Wholesale and retail trade	50.13
Food	49.19
Pulp, paper and paper-baesd manufacturing	46.49
Tobacco	45.56
Printing and recorded media replication	45.52
Rubber and plastic	45.41
Petroleum	44.7
Mining	44.02
Semiconductor	43.28
Clothing, Clothing accessory and fur clothing	43.08
Automobile and trailer	42.11
Transportation	42.06
Chemicals and chemical products (Except pharmaceuticals)	40.6
Furniture	40.08
Rubber and plastic	39.01
Service	38.07
Finance and insurance	35.75
Shipbuilding	35.00
Construction	25.5

연구 1을 통하여 가장 우수한 예측력을 보인 'TAN + CART + SVM' 모델을 가지고 주식 유동성 예측을 하였다. 실험 결과 전자장비 제조업, 가죽, 가방 및 신발 제조업, 목재 및 나무제품 제조업, 도매 및 소매업이 평균보다 좋은 예측치를 보였다. 그러나 건설업, 조선 산업, 금융 및 보험업, 서비스업, 섬유제품 제조업에 대해서는 'TAN + CART + SVM' 모델이 평균보다 낮은 예측치를 보였다. 본 결과에서는 Amihud 측정치를 5분위로 나누었기 때문에 60%정도 밖의 예측력을 보이지 못하였지만, '높음', '중간', '낮음' 등의 3분위로 나눈다면 예측력은 더 높아질 것이다.

산업별로 주식 유동성 예측에 차이가 있는 이유는 본 연구에서는 재무지표를 사용하여 주식 유동성을 예측하였기에 재무구조가 안정적인 산업의 경우 주식 유동성 예측 모델이 기업의 재무구조를 반영하는 예측모델을 만들 수 있었다. 그러나 조선 산업이나 건설업의 경우 수주 규모 및 공사 계약 한 건의 규모가 다른 산업들에 비해 크며, 이에 따른 수익성 및 재무지표 변화가 크기에[36] 주식 유동성 예측이 어렵다. 또한 서비스업, 금융 및 보험업은 글로벌 금융 위기 이후 회사 간 인수, 합병, 퇴출이 지속적으로 발생함으로서[37] 주식 유동성 예측 모델에 변수들이 왜곡되어 주식 유동성 예측이 어렵다. 이에 비하여 전자장비 제조업 등은 재무구조가 안정적이기에 주식 유동성 예측이 용이하다.

투자자와 회사 경영진 등 실무자에게 본 연구를 통해 얻은 결론은 크게 두 가지 의의를 갖는다. 첫째, 실무자들은 정확한 주식 유동성 예측을 통해, 서론에서 언급한 투자수익 실현[3] 및 회사 실적 향상[4]이 가능하다. 따라서 본 연구에서 확인한 데이터마이닝 방법은 실무자들의 의사결정에 도움을 줄 수 있다는 점에서 의의가 있다. 둘째, 주식시장에서는 실제 의사결정에 유용한 정보는 얻기 힘든 경우가 많다. 그러나 본 연구에서는 기업에서 일반 투자자 모두에게 공시하는 자료를 가지고 주식 유동성 예측을 하였다. 따라서 본 연구에서 사용한 예측 모델은 보다 적은 시간과 노력을 가지고 가치 있는 정보를 얻는다는 점에서 그 의의가 있다.

주식 유동성 및 이와 관련한 의사결정을 연구하는 연구자에게 본 연구를 통해 얻은 결론은 크게 세 가지 의의를 갖는다. 첫째, 복잡다단한 현실세계 및 주식시장에서 데이터마이닝 방법은 주식 유동성 예측을 가능하게 하고, 또한 예측력이 상대적으로 높다는 점에서 그 의의가 있다. 둘째, 다양한 데이터마이닝 방법을 합성하는 앙상블 방법은 주식시장의 다차원적이고 복잡한 현실을 반영한다는 점에서 의의가 있다. 셋째, 산업별 서로 상이한 재무적 특성을 모두 반영하는 하나의 알고리즘은 확인하기 힘들다는 한계 속에서 적어도 특정 산업에서는 더 좋은 예측력을 보이는 데이터마이닝 방법을 확인한다는 점에서 본 연구의 의의이다.

Ⅶ. Conclusion

본 연구는 투자자와 회사 경영진의 의사결정에 도움을 줄 수 있는 주식 유동성 예측 모델을 개발하는 데에 목적이 있다. 따라서 기존 기업의 재무상태 관련 변수와 주식 유동성간의 관계를 보는 데에서 한발 더 나아가 미래의 기업 주식 유동성 상태 예측을 위한 모델을 개발하였다. 실증 분석을 통하여 본 연구에서는 주식 유동성을 예측하기 위한 'TAN + CART + SVM' 앙상블 기법을 구축할 수 있었다. 이 모델은 기존 회귀분석 방법의 낮은 예측력을 극복함으로서 기업의 주식 유동성 예측을 할 수 있었다. 또한 전자장비 제조업, 가죽, 가방 및 신발 제조업, 목재 및 나무제품 제조업, 도매 및 소매업에서 'TAN + CART + SVM'모델이 주식 유동성에 관한 예측력이 높음을 확인할 수 있었다. 따라서 본 연구를 통해 투자자와 회사 경영진들은 미래에 변화할 기업의 주식 유동성 정도를 파악하여 의사결정에 도움을 받을 수 있다.

본 연구의 한계로는 우선 각각의 데이터마이닝 방법에 대한 매개변수를 디폴트 값으로 설정한 것에 있다. 본 연구에서 사용된 데이터마이닝 방법들은 매개변수를 어떻게 설정하느냐에 따라 예측력 결과가 달라질 수 있다. 그러나 본 연구에서는 주식 유동성을 파악 하는 데 기존에 사용되지 않았던 데이터마이닝 방법을 사용하는 것이 주요 목적이었기 때문에 매개변수에 따른 예측력 변화를 다루지 않았다. 그리고 본 연구를 통해 구축된 모델이 우리나라의 주요 산업인 건설업, 조선업 및 금융 및 보험업, 서비스업, 섬유제품 제조업의 주식 유동성 예측력이 떨어진다는 것이다. 그 이유로는 첫째, 산업 특성상 재무구조가 불안정하기 때문에 재무지표

를 반영한 주식 유동성 예측 모델이 제대로 구축되지 못하였기 때문이고 둘째, 건설업과 조선업의 주식 시장에 상장된 회사의 수가 적어 불규칙적인 시장상황에 소수의 기업이 더 큰 변동성을 보이기에 주식 유동성 예측을 제대로 구축할 수 없기 때문이다.

이러한 문제점을 극복하는 향후 연구의 진행 방향으로는 첫째, 본 연구에서 사용된 데이터마이닝 방법의 매개변수를 비교분석하여 최적의 알고리즘을 선택한다. 둘째, 본 연구에서 사용된 재무관련 변수 이외에 더 많은 변수들을 추가하여 기업의 주식 유동성 예측을 하는 것이다. 이 때 마코브 블랭킷(Markov Blanket)과 최고우선법칙(Best-First) 등의 데이터마이닝 방식을 이용하면 간결하면서도 주식 유동성 예측력이 높은 모델을 만들 수 있다. 셋째, 본 연구에서 사용된 우리나라의 상황 이외에 전 세계의 기업을 대상으로 훈련 자료(training data)을 만든 후 우리나라 기업에 대해 검증 자료(test data)을 구성하여 모델을 검증할 수 있다. 이 경우 나라 간 재무구조 차이는 크지 않음을 가정해야 한다.

본 연구는 우리나라 주식시장 자료를 이용하여 앙상블 방법 등 데이터마이닝 기법을 통한 기업의 주식 유동성을 예측하는 최초의 연구이다. 따라서 앞으로 기업 주식 유동성과 관련한 다양한 연구가 진행될 수 있는 계기가 될 것으로 기대한다. 또한 경영자와 투자자는 본 연구 결과를 토대로 관심 기업의 주식 유동성을 예측하여 이와 관련한 다양한 의사결정을 하는 데에 있어서 큰 도움을 얻을 것으로 기대한다.

REFERENCES

[1] H. C. Lee, "The Relation between Asset Liquidity and Stock Liquidity," Korean Journal of Business Administration, Vol. 27, No. 10, pp. 1691-1710, 2014.

[2] Korea Capital Market Institute, "Outlook for Korea`s stock and bond markets," Seoul, S. W. Hwang and S. H. Kang, 2015.

[3] K. Mazouz, W. Daya and S. Yin, "Index revisions, systematic liquidity risk and the cost of equity capital," Journal of International Financial Markets, Institutions and Money, Vol. 33, pp. 283-298, 2014.

[4] M. L. Lipson and M. Sandra, "Liquidity and capital structure," Journal of Financial Markets, Vol. 12, No. 4, pp. 611-644, 2009.

[5] H. J. Ko, Y. S. Park and H. S. Lee, " The Empirical Analysis on the Relation between Volatility of Liquidity and Return," Korean Journal of Business Administration, Vol. 22, No. 5, pp. 2873-2893, 2009.

[6] Y. Amihud, and H. Mendelson, “Liquidity and stock returns,” Financial Analysts Journal, Vol. 42, No. 3, pp. 43-48, 1986.

[7] A. S. Turnbull, R. W. White and B. F. Smith, " In search of liquidity: The block broker's choice of where to trade cross-listed stocks," Journal of Economics and Business, Vol. 62 No. 1, pp. 20-34, 2010.

[8] L. Kryzanowski and S. Lazrak, "Liquidity minimization and cross-listing choice: Evidence based on Canadian shares cross-listed on U.S. venues," Journal of International Financial Markets, Institutions and Money, Vol. 19, No. 3, pp. 550-564, 2009.

[9] R. Gopalan, O. Kadan and M. Pevzner, "Asset liquidity and stock liquidity," Journal of Financial and Quantitative Analysis, Vol. 47, No. 2, pp. 333-364, 2012.

[10] K. S. Cho, H. C. Shin, "A Study on the Effects of Block Ownership on Trading Activity and Market Liquidity in Korean Stock Market," Korean Journal of Business Administration, Vol. 26, No. 1, pp. 131-148, 2013.

[11] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference," Morgan Kaufmann, 1988.

[12] B. Yet, K. Bastani, H. Raharjo, S. Lifvergren, W. Marsh and B. Bergman, "Decision support system for Warfarin therapy management using Bayesian networks," Decision Support Systems, Vol. 55, No. 2, pp. 488-498, 2013.

[13] Y. Zuo and E. Kita, "Stock price forecast using Bayesian network," Expert Systems with Applications, Vol. 39, No. 8, pp. 6729-6737, 2012.

[14] F. Zheng, G. I. Webb, P. Suraweera and L. Zhu, "Subsumption resolution: an efficient and effective technique for semi-naive Bayesian learning," Machine Learning, Vol. 87 No. 1, pp. 93-125, 2012.

[15] G. I. Webb, J. R. Boughton, F. Zheng and K. M. Ting, "Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification," Machine Learning, Vol. 86, No. 2, pp. 233-272, 2012.

[16] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," Expert Systems with Applications, Vol. 42, No. 6, pp. 2928-2934, 2015.

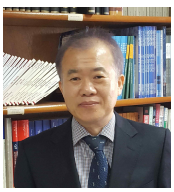
[17] L. Bouchaala, A. Masmoudi., F. Gargouri. and A. Rebai, "Improving algorithms for structure learning in Bayesian Networks using a new implicit score," Expert System Application, Vol. 37, No. 7, pp. 5470-5475, 2010.

[18] R. O. Duda, P. E. Hart. and D. G. Stork, "Pattern classification," Journal of Classification, Vol. 24, No. 2, pp. 305-307, 2007.

[19] J. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufman, 1993.
[20] S. Lee, "Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls," Decision Support Systems, Vol. 49, No. 4, pp. 486-497, 2013.
[21] L. Rutkowski, M. Jaworski, L. Pietruczuk and P. Duda, "The CART decision tree for mining data streams," Information Sciences, Vol. 266, No. 10, pp. 1-15, 2014.
[22] Y. Lin, H. Guo. and J. Hu, "An SVM-based Approach for Stock Market Trend Prediction," Proceedings of International Joint Conference on Neural Networks, pp. 1-7, 2013.
[23] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural processing letters, Vol. 9, No. 3, pp. 293-300, 1999.
[24] L. Zhou, K. K. Lai and L. Yu, "Least squares support vector machines ensemble models for credit scoring," Expert Systems with Applications, Vol. 37, No. 1, pp. 127-133, 2010.
[25] M. T. Hagan, H. B. Demuth and M. H. Beale, "Neural network design", Boston: Pws Pub, 1996.
[26] H. C. W. Lau, G. T. S. Ho and Y. Zhao, "A demand forecast model using a combination of surrogate data analysis and optimal neural network approach," Decision Support Systems, Vol. 54, No. 3, pp. 1404-1416, 2013.
[27] P. Hájek, "Municipal credit rating modelling by neural networks," Decision Support Systems, Vol. 51, No. 1, pp. 108-118, 2011.
[28] T. G. Dietterich, "Ensemble learning," The handbook of brain theory and neural networks, Vol. 2, pp. 110-125, 2002.
[29] K. C. Lee and K. Choi, "A study on the classification properties of firms to be subject to accounting disclosure reviews and investigations: Comparison of Bayesian Network, C5.0, and ensemble prediction methods," Korean Management Review, Vol. 36, No. 3, pp. 705-737, 2007.
[30] L. I. Kuncheva and J. J. Rodríguez, "Classifier ensembles for fMRI data analysis: an experiment," Magnetic Resonance Imaging, Vol. 28, No. 4, pp. 583-593, 2010.
[31] E. Fersini, E. Messina and F. A. Pozzi, "Sentiment analysis: Bayesian Ensemble Learning," Decision Support Systems, Vol. 68, 26-38, 2014.
[32] J. K. Bae, "An integrated approach to predict corporate bankruptcy with voting algorithms and neural networks," Korean Business Review, Vol. 3, No. 2, pp. 79-101, 2010.
[33] C. W. Yang, "Comparisons of Liquidity Measures in the Korean Stock Market," Asian Review of Financial Research, Vol. 25, No. 1, pp. 37-88, 2012.
[34] P. M. Dechow, R. G. Sloan and A. P. Sweeney, "Detecting earnings management," the Accounting Review, Vol. 70, No. 2, pp. 193-225, 1995.
[35] J. Han, M. Kamber and J. Pei, "Data mining. concepts and techniques," Morgan Kaufmann, 2012.
[36] K. S. Cho, S. H. Lee and J. J. Kim, "Influence of Overseas Construction Business on Construction Companies' Financial Stability," Korean journal of construction engineering and management, Vol. 14, No. 1, pp. 43-51, 2013.
[37] K. J. Kim and H. S Kim, "A Study on the Characteristics of Asymmetric Volatility by Industry in Korean Stock Market ", Korean Journal of Business Administration, Vol. 21, No. 6, pp. 2947-2964, 2008.

Authors

Eun Chan Bae is the candidate for B.S. in the Department of Global Business Administration, Sungkyunkwan University, Korea. He is interested in data-mining and artificial intelligence.



Kun Chang Lee is a full professor of MIS in SKK Business School at Sungkyunkwan University. He is now in charge of Creativity Science Research Institute (CSRI) and Health Mining Research Center (HMRC) as well, Sungkyunkwan University. His recent research interests lie in data mining, health informatics, creativity science, Human-Robot Interaction (HRI), and artificial intelligence techniques in decision making analysis.