Journal of The Korea Society of Computer and Information
Vol. 22 No. 11, pp. 111-116, November 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.11.111

Predicting stock price direction by using data mining methods : Emphasis on comparing single classifiers and ensemble classifiers

Kyun Sun Eo*, Kun Chang Lee**

Abstract

This paper proposes a data mining approach to predicting stock price direction. Stock market fluctuates due to many factors. Therefore, predicting stock price direction has become an important issue in the field of stock market analysis. However, in literature, there are few studies applying data mining approaches to predicting the stock price direction. To contribute to literature, this paper proposes comparing single classifiers and ensemble classifiers. Single classifiers include logistic regression, decision tree, neural network, and support vector machine. Ensemble classifiers we consider are adaboost, random forest, bagging, stacking, and vote. For the sake of experiments, we garnered dataset from Korea Stock Exchange (KRX) ranging from 2008 to 2015. Data mining experiments using WEKA revealed that random forest, one of ensemble classifiers, shows best results in terms of metrics such as AUC (area under the ROC curve) and accuracy.

▸Keyword: Stock price direction prediction, Data Mining, Feature selection, Single classifiers, Ensemble classifiers

I. Introduction

본 연구는 기업의 재무자료와 주식가격정보를 활용하여 미래의 주가방향을 예측한다. 주가방향 예측 (Stock price direction prediction)이란 주가가 전년도 대비 일정 비율 이상 오를지 또는 내릴지를 예측하는 것이다. 본 연구는 데이터마이닝 기법(Data mining)을 통하여 주가에 영향을 미치는 특성(Features)을 확인하고, 이를 바탕으로 주가방향 예측을 위한 적정 데이터마이닝 분류기(Classifier)를 제시하고자 한다.

금융업계에서 주가방향을 예측하는 것은 중요한 주제이다[1-3]. 주가방향 예측에 관한 선행연구의 대부분은 데이터마이닝 기법을 이용하고 있다[4-6]. 해당 연구에서는 다양한 분류기의 주가방향 예측 성능을 비교하고 또한 특성추출의 성과를 비교하고 있다[7-8].

반면 본 연구에서는 특성추출을 실행하고 앙상블 분류기를 포함한 다양한 분류기 간의 성능을 비교하고자 한다. 본 연구에서 사용하는 분류기는 다음과 같다. 첫째, 단일분류기(Single classifier)로서 인공신경망(Neural Network: NN), 의사결정트리(Decision Tree: DT), 서포트 벡터 머신(Support Vector Machine: SVM)과 로지스틱 회귀 (Logistic Regression: LR)를 이용한다. 둘째, 앙상블 분류기(Ensemble Classifier)로서 아다부스트(Adaboost: AB), 랜덤포레스트(Random Forest: RF), 배깅(Bagging: BA), 그리고 스태킹(Stacking: ST)을 적용하였다.

본 연구의 연구질문 (RQ: Research Question)은 다음과 같다.

RQ1: 데이터 마이닝 기법을 이용하여 주가방향 예측에 필요한 특성추출(Feature Selection, FS)을 시도한다.

RQ2: RQ1에서 추출된 특성을 이용하여 주가방향 예측을 위한 적정 데이터마이닝 분류기를 제시한다.

• First Author: Kyun Sun Eo, Corresponding Author: Kun Chang Lee
*Kyun Sun Eo (eokyunsun@gmail.com), MS student in MIS, Sungkyunkwan University, Seoul 03063, Republic of Korea
**Kun Chang Lee (kunchanglee@gmail.com), Professor at SKK Business School/SAIHST (Samsung Advanced Institute of Health Sciences & Technology), Sungkyunkwan University, Seoul 03063, Republic of Korea
• Received: 2017. 03. 19, Revised: 2017. 05. 19, Accepted: 2017. 08. 09.

본 연구의 구성은 다음과 같다. 2장에서는 주가방향 예측에 관한 선행연구를 간략하게 소개하고, 데이터마이닝 기법 중 특성추출과 분류기에 대해 설명했다. 3장에서는 본 연구에 사용된 데이터와 연구 방법을 설명했다. 4장에서는 주가방향 예측에 필요한 특성과 적정 데이터마이닝 분류기의 성과를 비교했다. 마지막으로 5장에서는 본 연구의 시사점과 향후 연구방향에 대해 설명했다.

II. Preliminaries

2.1 Previous Studies

효율적 시장가설(EMH)[9]은 모든 정보가 즉각 공유되어 금융자산의 가격이 충분히 반영되는 시장을 말한다. 투자자가 가진 정보는 이미 주가에 반영이 되기 때문에 투자자가 가진 정보로는 초과 수익을 얻을 수 없다. 하지만 선행연구는 데이터마이닝 방법을 이용해 과거의 데이터를 분석해 예측 하고 있다. 연구자들은 복잡하고 비선형(non-linear)인 주식 시장을 분석해 더 나은 결과를 만들기 위해 노력하고 있고, 이에 필요한 기술을 고안하기 위해 연구 하고 있다[10-11].

주가방향을 예측 하는 방법은 대표적으로 3가지가 있다. 첫째 기술적 분석, 둘째 시계열 예측, 그리고 데이터 마이닝 방법[12] 등이 있는데 예측 방법은 다음과 같다. (1)기술적 분석방법은 차트를 이용한 도구로써 분석가들이 구매 또는 판매에 관한 결정을 할 때 사용된다. (2)시계열 예측 방법은 과거의 주가를 분석하여 미래의 주가를 예측하는 것이다. 시계열 예측방법으로는 자기회귀모형(autoregressive method), 이동평균법(moving average model), 자기회귀-이동평균(autoregressive-moving average model) 등이 있다. (3)데이터마이닝 방법은 복잡하고 대용량의 데이터에서 가치 있는 정보를 추출하는 것이다[13]. 주가방향을 예측할 때 생기는 문제점 중 하나는 방대한 양의 데이터를 처리하는 것이다. 데이터마이닝 방법을 사용하지 않고 대용량의 데이터를 분석하는 것은 데이터의 양이 매우 많기 때문에 데이터의 숨은 정보를 정확히 파악할 수 없다.

주가방향 예측에 대한 선행연구는 Table 1과 같다. 선행연구는 로지스틱 회귀분석, 인공신경망과 서포트 벡터머신을 이용해 주가방향을 예측 했고, 앙상블 분류기인 랜덤포레스트, 아다부스트 등을 사용했다[6]. 선행연구는 서포트 벡터 머신과 랜덤포레스트가 높은 성능을 보인다고 제시했다[5][7]. 그리고 아다부스트는 랜덤포레스트 만큼 높은 예측력을 가지고 있다[4]. 본 연구는 WEKA를 이용하여 Table 1에 표시한 단일분류기와 앙상블 분류기를 이용해 분석을 진행했다.

2.2 Feature Selection

Feature Selection은 특성추출을 의미한다. 이는 분류기법을 실행하기 전에 불필요한 특성을 배제하고 실제 타겟 변수에 영향을 주는 특성들을 추출하는 것이다. 특성추출을 통해 분류기의 정확도를 향상시키고 데이터의 본질적인 특징을 파악할 수 있다[5]. 특성추출은 필터기반 추출 방법(Filter-based method)과 래퍼방법(Wrapper-based method)이 주로 사용된다[13]. 래퍼방법은 성능이 좋은 방법 중 하나이기 때문에 인기 있는 방법이다[14]. 본 연구에서는 유통업의 주가방향 예측에 필요한 특성을 파악하고자 다음과 같은 특성추출 방법을 사용했다.

본 연구는 상관관계 기반에 의한 특성추출(Correlation based Feature Selection) 방법과 특성추출을 위한 검색기능인 Best-first 방법을 사용했다.

2.3 Data Mining

2.3.1 Single Classifier Decision tree

의사결정트리는 복잡하고 대용량의 자료를 처리할 때 특성의 성질이나 특성 간의 관계를 쉽게 판별할 수 있는 장점을 가

Table 1. Classifiers for stock price direction prediction used in literature.

Authors	Year	Feature Selection	Single classifier				Ensemble classifier			
			DT	LR	NN	SVM	RF	AB	BA	ST
P. N. Rodriguez & A. Rodriguez. [4]	2004		√	√	√		√	√		
M. Kumar & M. Thenmozhi. [5]	2006			√	√	√	√			
Y. Kara et al. [10]	2011				√	√				
L. P. Ni et al. [8]	2011	Information Gain, Symmetrical uncertainty, ReliefF, Cfs, OneR, Improved Fractal				√				
J. Patel et al. [7]	2015				√	√	√			
M. Ballings et al. [6]	2015			√	√	√	√	√		
Y. Shynkevich et al. [11]	2016	Chi-square Bi-normal Separation			√	√				
This study		Cfs	√	√	√	√	√	√	√	√

지고 있다. 의사결정트리는 좋은 성능을 가지기 때문에 인기 있는 분류기 중의 하나이다[15-16]. 본 연구는 WEKA에서 제공하는 J4.8분류기를 사용 했다. WEKA에서 사용하고 있는 J4.8 분류기은 C4.5분류기를 기반으로 개발된 분류기이다[17].

Neural Network

인공신경망 방법은 인간 뇌의 기본구조를 이루는 뉴런(neuron)의 모형을 기반으로 만든 모델이다. 인공신경망은 복잡한 비선형 함수(non-linear function)를 구축할 수 있다[15]. 인공신경망은 입력층, 은닉층, 출력층으로 이루어져 있으며 은닉층에 존재하는 각각의 노드들은 의사결정트리의 잎 노드와 동일한 역할을 한다[16]. 인공신경망은 신경망 구조가 주어졌을 때 역전파 알고리즘(Back-propagation)을 통해 가중치를 결정 한다.

Logistic Regression

로지스틱 회귀분석은 종속변수가 이항형(Binominal) 또는 다항형(Multinominal)일 때 사용하는 회귀분석 방법이다[18]. 본 연구는 종속변수가 범주형 데이터이기 때문에 분류 기법이다.

Support Vector Machine

서포트 벡터 머신은 선형 모형과 인스턴스가 기반인 학습 알고리즘이 합쳐진 지도 학습 모형이다[14]. 서포트 벡터 머신은 데이터를 선형이나 비선형으로 구분하는데 효과적인 분류기이다. 비선형 분류인 경우에는 비선형 커널 함수를 이용하여 입력값을 초평면으로 변환시킨다[6]. 본 연구는 다항커널함수를 이용하였다.

2.3.2 Ensemble Classifier

앙상블 분류기는 여러 개의 분류기를 조합해 더 나은 성능을 내는 분류기이다.[6]. 앙상블 분류기는 동일한 타입의 분류기를 재구성하는 배깅, 부스팅 방법이 있고, 다양한 분류기를 통해 구성된 모델에 적용하는 스태킹이 있다.

III. The Proposed Scheme

3.1 Data and Variables

본 연구의 데이터는 NICE 평가 정보(주)에서 제공하는 KIS-value에서 수집했다. 2008년부터 2015년도 까지 근속된 기업으로 KOSPI에 상장된 유통업이다. 본 연구에서 사용한 데이터는 다음과 같다. 독립변수 개수는 109개이며, 변수의 내용은 Table 2와 같다. 레코드 개수는 1184개, 회사의 수는 37개이다. 2008년부터 2015년까지 분기별 재무자료와 주식종가로 구성하였다.

본 연구의 목적을 위해서 타겟 변수를 다음과 같이 설정했다. 타겟 변수는 월별 주식종가를 분기별 평균으로 구성한 다음에 2007년도의 주식종가와 다음연도인 2008년도를 비교했다. 2009년도도 2008년도와 같이 전년도인 2008년도와 비교하여 2015년도 까지 비교했다. 그리고 전 년도 대비 주식종가의 성장률을 임계점(Threshold)을 지정해 비교했다. 타겟 변수별 임계점은 다음과 같다. TH1은 25%, TH2는 35%, 15%, TH3은 35%, 25% 15%, TH4는 35%, 25%, 15%, 5%이다.

Table 2. Description of Features

No	feature
1	Total assets growth
2	Tangible assets growth
3	Current assets growth
4	Inventories growth
5	Shareholder's equity growth
6	Net sales growth
7	Operating income growth
8	Income before income tax expense growth
9	Net income growth
10	No. of employee growth
11	Operating income to total assets
12	Income before income tax expense to total assets
13	Net income to total assets
14	Income before income tax expense
15	Net income before financial expenses to avg. total assets
16	Operating income to operating capital
17	Income before income tax expense to equity
18	Net income to shareholder's equity
19	Income before income tax expense to capital stock
20	Net income to capital stock
21	Income before income tax expense to net sale
22	Net income to net sales
23	Gross profits to net sales
24	Operating income to net sales
25	Total expenses to total revenue
26	COGS to net sales
27	Depreciation ratio
28	Depreciation/total cost
29	Personnel expenses/total cost
30	Taxes / Income before income taxes
31	Taxes / total cost
32	Financial expenses / total liabilities
33	Financial expenses / total borrowings
34	Financial expenses / total expenses
35	Financial expenses / net sales
36	Times interest earned-operating act. basis
37	Times interest earned-operating income basis
38	Times interest earned-ordinary income basis
39	Times interest earned-income before income taxes basis
40	Dividend ratio
41	Dividend to net income
42	Coverage ratio
43	Debt coverage ratio
44	Loan efficiency ratio
45	EBIT/net sales
46	EBITDA/net sales
47	EBITDA/financial expenses
48	Equity to total assets
49	Current ratio
50	Quick ratio
51	Cash ratio
52	Non current assets ratio
53	Non current assets to equity & LT liabilities
54	Total liabilities to shareholder's equity

55	Current liabilities to shareholder's equity
56	Non-Current liabilities to shareholder's equity
57	Total borrowings to total assets
58	Total borrowings to shareholder's equity
59	Total borrowings/net sales
60	A/R to trade account payable
61	A/R to merchandise & finished goods
62	Trade account payable to inventories
63	Inventories to NWC
64	Non-Current liabilities to NWC
65	NWC to total assets
66	Reserves ratio
67	Reserves to total disposal amount of R/E
68	R/E to total assets
69	R/E to paid-in capital
70	Total CF to total liabilities
71	Total CF to total borrowings
72	Total C/F to total assets
73	Total C/F to net sales
74	Net CF to total borrowings
75	Total assets turnover
76	Equity turnover
77	Paid-in capital turnover
78	NWC turnover
79	Operating capital turnover
80	Non-Current assets turnover
81	Tangible assets turnover
82	Inventories turnover 1
83	Merchandise & finished goods turnover
84	Raw materials turnover
85	WIP turnover
86	A/R turnover
87	Trade account payable turnover
88	Inventories turnover 2
89	Net operating capital turnover
90	Value-added per employee
91	Net sales per employee
92	Income before income tax expense per employee
93	Net income per employee
94	Personnel expenses per employee
95	Avg. tangible assets, net of CIP per employee
96	Machinery & equipment per employee
97	Total assets per employee
98	Efficiency of investment-avg. total assets
99	Efficiency of investment-avg. tangible assets, net of CIP
100	Efficiency of investment-avg. machinery
101	Value added to net sales
102	Labor cost to value added
103	Income before income taxes to value added
104	Personnel expenses to value added
105	Financial expenses to value added
106	Rent to value added
107	Taxes & dues to value added
108	Depreciation to value added
109	Closing price

3.2 Method

3.2.1 Feature Selection (FS)

특성추출은 분류기법을 실행하기 전에 불필요한 특성을 배제하고 실제 분류 성능에 영향을 주는 특성들을 추출하는 것이다. 특성추출은 분류기의 정확도를 향상시키고 데이터의 본질적인 특징을 파악할 수 있다[5].

3.2.2 Model Evaluation Criteria

본 연구는 AUC(Area under the Receiver operating Characteristics)를 적용했다[19]. AUC는 신호탐지 이론에서 적중확률(X축, True Positives, Sensitivity), 오경보확률(Y축, False Positive, 1-Specificity)을 나타내는 그래프이다. AUC는 0.5부터 1까지 예측확률을 나타내고 1에 가까울수록 높은 예측확률을 가진다. TP는 True Positives, TN은 True Negatives, P는 positives(event), N은 Negatives(non-event)이다.

3.2.3 Cross validation

본 연구는 cross validation을 이용하여 데이터마이닝 분류기를 검증했다[20]. 본 연구에서는 10-fold cross validation 방법을 이용했다. 10-fold cross validation방법은 데이터를 10개의 폴더로 나눈 후 9개 폴더는 분류기를 학습하기 위해 사용되고 남은 1개 폴더는 검증하기 위해 사용하는 하는 검증방법이다.

IV. Results

4.1 Results for RQ1

본 연구에서는 데이터마이닝 기법을 이용하여 주가방향 예측에 필요한 특성을 추출했다. 109개의 변수 중에서 특성추출을 통해 추출한 변수는 Table 3과 같다. TH1은 23개, TH2는 26개, TH3은 20개, TH4는 19개로 대폭 감소했다.

4.2 Results for RQ2

TH 별로 분류기의 AUC는 Table 4와 같다. AUC가 가장 높은 분류기는 랜덤포레스트이고 그다음으로 높은 것은 배깅으로 확인했다. TH별로 분류기의 정확도를 검증한 결과는 Table 5와 같다. 랜덤포레스트가 가장 높은 예측력을 보였고, 다음으로 배깅의 예측력이 높았다. 본 연구를 통해 분류기의 성능을 비교해본 결과 단일분류기보다 앙상블 분류기가 더 높은 성능을 가지는 것으로 확인했다. 위 결과를 바탕으로 TH별로 AUC와 Accuracy의 F-Test를 실시하였고, 랜덤포레스트와 배깅을 단일분류기의 차이를 검증하기 위해 Tukey 검정을 실시했다. 결과는 Table 6과 같다.

Table 3. Features selected by FS

	Feature
TH1	2, 4, 5, 6, 11, 14, 16, 17, 18, 19, 20, 35, 39, 42, 45, 52, 56, 57, 59, 66, 68, 76, 109
TH2	2, 4, 11, 12, 14, 16, 17, 18, 19, 20, 35, 37, 39, 45, 52, 55, 56, 57, 58, 59, 65, 66, 68, 69, 95, 109
TH3	4, 11, 12, 16, 17, 19, 20, 35, 37, 38, 49, 52, 56, 57, 59, 66, 68, 69, 95, 109
TH4	11, 16, 17, 18, 19, 20, 21, 32, 35, 37, 39, 47, 57, 59, 66, 68, 69, 93, 109

Table 6. Result of F-Test & Tukey Test

	F-Test		Tukey Test					
		sig.	Comparison	t-value	sig.	Comparison	t-value	sig.
TH1	AUC	0.000	RF-DT	6.4446	0.000	BA-DT	4.0707	0.126
	Accuracy	0.000	RF-LR	7.5865	0.000	BA-LR	5.2126	0.000
			RF-NN	7.2294	0.000	BA-NN	4.8555	0.000
			RF-SVM	7.4064	0.000	BA-SVM	5.0325	0.000
TH2	AUC	0.000	RF-DT	6.8731	0.000	BA-DT	3.2588	0.141
	Accuracy	0.000	RF-LR	8.9854	0.000	BA-LR	5.3710	0.001
			RF-NN	7.5764	0.000	BA-NN	3.9621	0.033
			RF-SVM	8.5413	0.000	BA-SVM	4.9270	0.003
TH3	AUC	0.000	RF-DT	9.3487	0.000	BA-DT	5.2942	0.000
	Accuracy	0.000	RF-LR	9.1618	0.000	BA-LR	5.1072	0.000
			RF-NN	9.4219	0.000	BA-NN	5.3673	0.000
			RF-SVM	9.2487	0.000	BA-SVM	5.1942	0.000
TH4	AUC	0.000	RF-DT	10.2258	0.000	BA-DT	4.5839	0.000
	Accuracy	0.000	RF-LR	9.4271	0.000	BA-LR	3.7851	0.000
			RF-NN	8.7245	0.000	BA-NN	3.0825	0.000
			RF-SVM	10.1280	0.000	BA-SVM	4.4861	0.000

Table 4. Area under ROC

	DT	LR	NN	SVM	RF	AB	BA	ST
TH1	0.69	0.65	0.59	0.50	0.81	0.60	0.75	0.50
TH2	0.72	0.65	0.67	0.50	0.81	0.53	0.75	0.50
TH3	0.68	0.61	0.65	0.50	0.81	0.55	0.74	0.50
TH4	0.71	0.63	0.65	0.50	0.81	0.54	0.74	0.50

Table 5. Analysis Result of Accuracy

	# of Features	DT	LR	NN	SVM	RF	AB	BA	ST
TH1	23	71.62	70.48	70.84	70.66	78.07	71.54	75.69	70.75
TH2	26	65.72	63.61	65.08	64.05	72.59	64.93	68.98	63.96
TH3	20	63.78	63.96	63.70	63.88	73.13	65.02	69.07	63.97
TH4	19	54.97	55.77	56.47	55.07	65.19	55.07	59.55	55.07

Table 6에 따르면 모든 TH에 대해 AUC와 Accuracy는 모두 통계적으로 유의하게 나왔다. 따라서 모든 TH에서 랜덤포레스트와 배깅의 경우 단일분류기와 비교하여 모두 통계적으로 유의함을 보였다.

V. Conclusions

본 연구는 주가방향 예측을 위해 단일분류기와 앙상블 분류기의 성능을 비교하였다. 본 연구에서 사용한 자료는 코스피에 상장된 유통업종 기업의 재무비율 자료와 주가자료이다. 해당 자료를 분류기로 분석하기 전 상관관계에 기반한 특성추출을 하였고, 그 결과 Table 3에서와 같은 특성을 추출하였다. 추출한 특성을 이용해 데이터마이닝 분류기의 성능을 비교한 결과, 랜덤포레스트와 배깅의 성능이 상대적으로 가장 우수하였다. 이같은 결과는 주가방향 예측시 랜덤포레스트의 성능이 우수하다는 선행연구의 결과와 일치한다[5-7]. 본 연구에서는 특성추출을 통해 일반 적인 업종이 아닌 특정 업종인 유통업 분석에 필요한 특성을 파악할 수 있었다. 또한 109개의 특성 중 26개 ~ 19 개로 대폭 감소시킴으로서 주가방향의 예측력을 높이고 주가방향에 필요한 특성을 제시했다. 향후 연구에서는 특성추출에 관한 새로운 방법을 개발하여 주가방향 예측을 수행할 수 있을 것이다.

REFERENCES

[1] R. Al-Hmouz, W. Pedrycz, & A. Balamash, "Description and prediction of time series: A general framework of granular computing". Expert Systems with Applications, Vol. 42, pp. 4830-4839, 2015

[2] S. Barak, & M. Modarres, "Developing an approach to evaluate stocks by forecasting effective features with data mining methods", Expert Systems with Applications, Vol. 42, pp. 1325-1339, 2015

[3] A. Booth, E. Gerding, & F. McGroarty, "Automated trading with performance weighted random forests and seasonality", Expert Systems with Applications, Vol. 41, pp. 3651-3661, 2014

[4] P. N. Rodriguez, & A. Rodriguez, "Predicting stock market indices movements", WIT Transactions on Modelling and Simulation, Vol.38, 2004.

[5] M. Kumar, & M. Thenmozhi, "Forecasting Stock index movement: A comparison of support vector machines and random forest". SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 24, 2006.

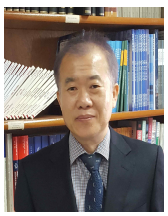
[6] M. Ballings, Dirk Van den Poel, Nathalie Hespeels, Ruben Gryp. "Evaluating multiple classifiers for stock price direction prediction", Expert Systems with Applications Vol. 42 pp. 7046-7056, 2015

[7] J. Patel, S. Shah, P. Thakkar, K. Kotecha, "Predicting

stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques", Expert Systems with Applications, Vol. 42, pp. 259-268, 2015
[8] L. P. Ni, Z. W. Ni, & Y. Z. Gao, "Stock trend prediction based on fractal feature selection and support vector machine", Expert Systems with Applications, Vol. 38(5), pp. 5569-5576, 2011
[9] B. G. Malkiel, & E. F. Fama, "Efficient capital markets: A review of theory and empirical work", The Journal of Finance, Vol. 25(2), pp. 383-417, 1970
[10] Y. Kara, M. A. Boyacioglu, & Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange", Expert systems with Applications, Vol. 38(5), pp. 5311-5319, 2011
[11] Y. Shynkevich, T. M. McGinnity, S. A. Coleman, & A. Belatreche, "Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning", Decision Support Systems, Vol. 85, pp. 74-83, 2016.
[12] T. Hellström, K. Holmströmm, "Predictable Patterns in Stock Returns". Technical Report Series IMa-TOM-1997-09, (August 9, 1998)
[13] Z. Li, W. Xu, L. Zhang, & R. Y. Lau, "An ontology-based Web mining method for unemployment rate prediction". Decision Support Systems, Vol. 66, pp. 114-122, 2014.
[14] R. Kohavi, & G. H. John, "Wrappers for feature subset selection". Artificial Intelligence, Vol. 97 , pp. 273-324 , 1997
[15] D. L. Olson, D. Delen, & Y. Meng, "Comparative analysis of data mining methods for bankruptcy prediction", Decision Support Systems, Vol. 52, No. 2, pp. 464-473, 2012.
[16] E. C. Bae, & K. C. Lee, "Predicting Stock Liquidity by Using Ensemble Data Mining Methods", Journal of The Korea Society of computer and Information, Vol. 21, No. 6, pp. 9-19, 2016.
[17] R. Quinlan, "C4.5: Programs for machine learning". San Mateo: Morgan Kaufmann Publishers, 1993
[18] A. Dag, A. Oztekin, A. Yucel, S. Bulur, F. M. Megahed, "Predicting heart transplantation outcomes through data analytics", Decision Support Systems Vol. 94 pp. 42-52, 2017
[19] F. Provost, T. Fawcett, & R. Kohavi, "The case against accuracy estimation for comparing induction algorithms", In Proceedings of the fifteenth international conference on machine learning (pp. 45-453), Morgan Kaufmann, 1997
[20] S. Arlot, & A. Celisse, "A survey of cross-validation procedures for model selection", Statistics Surveys, Vol. 4, pp. 40-79, 2010

Authors

Kyunsun Eo is a MS student in Business Administration at Sungky unkwan University, Korea. He is interested in data mining, machine learning, 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 interested in data mining, health informatics, creativity science, Human-Robot Interaction (HRI), and artificial intelligence techniques in decision making analysis.