doi:10.3772/j.issn.1006-6748.2023.02.002

GHM-FKNN: a generalized Heronian mean based fuzzy k-nearest neighbor classifier for the stock trend prediction①

WU Zhenfeng(吴振峰)*, WANG Mengmeng②**, LAN Tian*, ZHANG Anyuan***
(*Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)
(**School of Economics, Renmin University of China, Beijing 100872, P. R. China)
(***Shandong Provincial Center for Quality Control of Feed and Veterinary Drug, Jinan 250022, P. R. China)

Abstract

Stock trend prediction is a challenging problem because it involves many variables. Aiming at the problem that some existing machine learning techniques, such as random forest (RF), probabilistic random forest (PRF), k-nearest neighbor (KNN), and fuzzy KNN (FKNN), have difficulty in accurately predicting the stock trend (uptrend or downtrend) for a given date, a generalized Heronian mean (GHM) based FKNN predictor named GHM-FKNN was proposed. GHM-FKNN combines GHM aggregation function with the ideas of the classical FKNN approach. After evaluation, the comparison results elucidated that GHM-FKNN outperformed the other best existing methods RF, PRF, KNN and FKNN on independent test datasets corresponding to three stocks, namely AAPL, AMZN and NFLX. Compared with RF, PRF, KNN and FKNN, GHM-FKNN achieved the best performance with accuracy of 62.37% for AAPL, 58.25% for AMZN, and 64.10% for NFLX.

Key words: stock trend prediction, Heronian mean, fuzzy k-nearest neighbor (FKNN)

0 Introduction

The financial market plays an important role in the resource allocation and operation of the modern economy. In particular, the stock market and its trends are highly volatile in nature, which attracts researchers to capture the volatility and predict the future direction of stock price movements, whether it is an uptrend or a downtrend. Since the stock market generates a large amount of non-stationary time series data dominated by chaos every day, it becomes a challenging problem to forecast the future trend of stocks based on past stock data[1].

Although some theories such as the efficient market hypothesis[2] and the random walk hypothesis[3] stated that stock market prices are essentially unpredictable, many studies have elucidated that the stock trend could be partially predicted with the use of text mining and machine learning algorithms[4]. Technical and fundamental analysis are the two major approaches to predict the stock trend[5]. Technical analysis considers past price and volume to predict the future trend while fundamental analysis is mainly based on macroeconomic analysis, industry analysis and company analysis to get some insights[6]. To achieve high profits with low-risk stocks, investors have used technical and fundamental analysis to predict stock market price for investment decision making[4]. Therefore, accurate stock trend prediction is critical and fundamental to minimize risks and maximize profits from stocks.

Over the past few decades, several computational methods have been proposed to predict the future trend of a specific stock or overall market[7]. Recent advances in stock trend prediction mainly fall into four categories——statistical approach, pattern recognition, machine learning, and sentiment analysis.

Bhuriya et al.[8] implemented a statistical approach for predicting the Tata Consultancy Services stock price based on five features, namely open, volume, high, low, and close price. They compared the performances of regression model variants and reported that the linear regression (LR) model had a confidence value of 0.97, outperforming the polynomial and Radial Basis Function regression models. Kim et al.[9] constructed a pattern-matching trading system based on

① Supported by the National Key Research and Development Program (No. 2019YFA0707201) and the Key Work Program of Institute of Scientific and Technical Information of China (No. ZD2022-01, ZD2023-07).
② To whom correspondence should be addressed. E-mail: wmm0927@ruc.edu.cn.
Received on Aug. 8, 2022

a dynamic time-warping algorithm that identifies movement patterns in morning market data and determines afternoon liquidation strategy. Their approach can provide stable and efficient trading strategies with relatively low trading frequency. Khan et al.[10] implemented stock market prediction using different machine learning classifiers and social media, news. They reported consistent results using the random forest (RF) algorithm with 77.10% accuracy under 10-fold cross-validation test. Kalyani et al.[6] created three different stock trend prediction models using news sentiment analysis to explore the relationship between news and stock trends. The comparison results of RF, support vector machine, and Naive Bayes algorithms showed that RF was the best performing algorithm in all test cases, with accuracy between 88% and 92%. It is worthy of note that one of the latest research using news sentiment analysis and technical indicators implemented in big data computing platform-spark also reported RF was the best performing model with a 63.58% test accuracy compared with LR and gradient boosting machine[11]. Recent study also showed that the probabilistic random forest (PRF) outperforms RF in noisy datasets[12]. Although above methods have achieved acceptable performance, they are still far from being accurate.

In this study, to develop a more accurate model for predicting stock trends, a fuzzy k-nearest neighbor (FKNN) predictor based on generalized Heronian mean (GHM) called GHM-FKNN is proposed. GHM-FKNN determines the stock trends associated with a given date class labels (uptrend or downtrend) on the nearest local Heronian mean vector using the k-nearest neighbor (KNN) concept. The main algorithm of GHM-FKNN can overcome the class domination of the relative KNN by averaging all KNN vectors for each class to fully explain the class distribution. The latest datasets and nine features used in the recent study by Taylan et al. were collected and applied. Besides, 10-fold cross-validation and independent tests were used to evaluate model performance. GHM-FKNN is effective and improves the prediction accuracy of future stock trends. The comparison results showed that GHM-FKNN outperformed four existing predictors RF, PRF, KNN and FKNN on independent test datasets.

1 Methods

In this study, the key idea of GHM-FKNN is to make a binary classification prediction of whether a stock is in an uptrend or a downtrend on a given day based on some technical indicators and financial news information. The prediction process of GHM-FKNN is based on three main stages, namely data preprocessing, feature encoding, and model construction and evaluation. The framework of the GHM-FKNN method is displayed in Fig. 1.

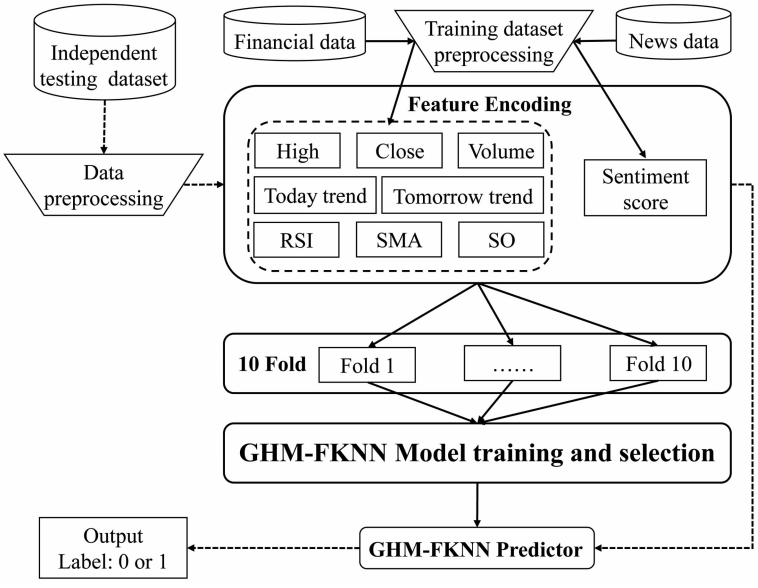

Fig. 1 The framework of GHM-FKNN

1.1 Data preprocessing

At this stage, the raw data consists of two parts—— financial data and financial news data. Financial data includes at least six features including date, the highest price of the day, the lowest price of the day, the

opening price, the closing price, and the trading volume. Financial data needs to be converted to spreadsheet format and missing values should be removed. Financial news data are in a text format written in English. Financial news data are preprocessed based on the following steps: (1) remove irregular characters with regular expressions, (2) convert all letters to lower cases, (3) remove stop words, and (4) convert each news text into word vector. All preprocessed data are fed into the feature encoding stage.

1.2 Feature encoding

Highly inspired by a recent research which has demonstrated the significance of technical indicators and technical analysis in predicting the stock market[11], the same following nine identical features for machine learning algorithms are also used.

(1) High

This feature refers to the highest price of a given day.

(2) Close

This feature refers to the closing price of a given day.

(3) Volume

This feature refers to the number of shares traded over the course of a given day.

(4) Today_trend

This feature reflects the today's trend (uptrend or downtrend) of a stock on a given date[10], which can be formulated as

$$\text{Today\_trend} = \begin{cases} 0 & P_c - P_o \geqslant 0 \\ 1 & P_c - P_o < 0 \end{cases} \tag{1}$$

where, P_c and P_o represent the closing price and the opening price of a given trading day, respectively. Today_trend $\in \{0, 1\}$, where uptrend is encoded by 0 and downtrend is encoded by 1.

(5) Tomorrow_trend

This feature reflects the tomorrow's trend (uptrend or downtrend) of a stock on a given date which can be formulated as

$$\text{Tomorrow\_trend} = \begin{cases} 0 & P_{tmc} - P_{tdc} \geqslant 0 \\ 1 & P_{tmc} - P_{tdc} < 0 \end{cases} \tag{2}$$

where, P_{tmc} and P_{tdc} represent the closing price of the next day and the closing price of the current day, respectively. Tomorrow_trend $\in \{0, 1\}$, where uptrend is encoded by 0 and downtrend is encoded by 1.

(6) Relative strength index (RSI)

This feature is a momentum indicator that evaluates overbought or oversold conditions by measuring the magnitude of recent price changes for various assets[13]. The description of RSI can be represented as

$$RSI = 100 - 100/(1 + \text{Avg}U/\text{Avg}D) \tag{3}$$

where Avg$U$ and Avg$D$ represent the average of all up and down moves in the last N price bars, respectively. N is the period of RSI. In this study, the period N is fixed at 14.

(7) Simple moving average (SMA)

This feature is one of the most commonly used technical indicators, referring to the average price of a stock over a set period of time[14]. The description of SMA can be written as

$$SMA = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{4}$$

where, P_i represents the closing price in the last N price bars. N is the period of the SMA. In this study, the period N is fixed at 14.

(8) Stochastic oscillator (SO)

This feature is a commonly used momentum indicator that compares a specific closing price of a security to its series of prices over a set time period[15]. SO is used to generate overbought and oversold trading signals, utilizing a bounded value range of 0 – 100 which can be denoted as

$$SO = \left(\frac{P_c - L_N}{H_N - L_N}\right) \times 100 \tag{5}$$

where, P_c represents the most recent closing price. L_N and H_N represent the lowest and highest price of the previous N trading sessions, respectively. In this study, the period N was fixed at 14.

(9) Sentiment score

Some of previous studies have shown that news polarity may influence changes in stock trends[6,11,14]. Highly inspired by previous research, the same sentiment analysis approach for calculating sentiment score are also adopted to analyze the preprocessed financial news data[11]. The calculated sentiment score for each news item ranges from -1 to 1. A news is positive if its sentiment score is close to 1 and negative if its sentiment score is close to -1. The news is neutral if its sentiment score is around 0.

1.3 Model construction and evaluation

1.3.1 GHM-FKNN algorithm

In principle, after encoding the features, any statistical machine learning algorithm can be applied to predict the stock trend. It is common practice to find a suitable classifier that can accurately identify whether a stock is in an uptrend or a downtrend. This paper mainly focuses on the improvement of KNN algorithm for its good performances reported in Ref. [16]. The KNN algorithm is a commonly used machine learning algorithm that clusters samples by calculating their distances[17]. The key idea of KNN is that if the majority of the KNN

of a sample belongs to a class based only on distance proximity, then the sample also belongs to a class. Since the classification performance of KNN is usually degraded on datasets due to the presence of outliers, a new algorithm GHM-FKNN based on GHM is proposed. GHM-FKNN combines the generalized Heronian mean aggregation function with the classical idea of the fuzzy KNN approach. The GHM-FKNN algorithm uses KNN concept to determine the class labels of unclassified samples based on the nearest local mean vector. Uptrend and downtrend are encoded by 0 and 1, respectively. The pseudo code of GHM-FKNN algorithm can be seen as Algorithm 1.

Algorithm 1 Pseudo code of GHM-FKNN

Input: X, the training data; C, the set of decision classes; y, the object to be classified
Output: Classification for y
Procedure Begin
1: Calculate the cosine distance $d(y, x_i)$, where $x_i \in X$ $(i = 1, 2, \cdots, n)$ and n is the number of samples in X.
2: Arrange the calculated n cosine distances in non-decreasing order.
3: Take the first K distances from this sorted list and find those K points corresponding to these K-distances.
4: Get the set of unique classes C_K among K points.
5: **if** length(C_K) = = 1 **then**
6: The class is the element of set C_K.
7: output Class
8: **else**
9: **for each** c in C_K **do**
10: Calculate the generalized Heronian mean $h(c, \text{closestPoints}, p, q)$.
11: Calculate the cosine distance $d(y, h)$.
12: Calculate the fuzzy membership values of an unclassified sample to all K nearest neighbors.
13: Arrange the fuzzy membership values in non-decreasing order. The class with the highest membership degree is the predicted class.
14: **end**
15: **output** Class
End

The main process of the proposed GHM-FKNN algorithm is as follows.

Firstly, KNNs are obtained by arranging the calculated cosine distances between the unclassified samples and training data. These KNNs are then grouped into the classes they belong to. The cosine distance is defined as

$$d(y, x_i) = 1 - \frac{y \cdot x_i}{\| y \| \cdot \| x_i \|} \tag{6}$$

where y is the unclassified sample and X is the training data with n samples; $x_i \in X = \{x_1, x_2, \cdots, x_n\}$.

The previous study has shown that the Bonferroni mean (BM) based fuzzy k-nearest centroid neighbor classifier is robust to outliers and can overcome class domination of its neighbors in datasets with class imbalance[18]. However, BM-based classifier have some drawbacks, which redundantly considers the interrelationship between two variables. Fortunately, the generalized Heronian mean, a powerful multi-criteria decision making aggregation function in an information fusion system[19], can deal with the interrelationship between two variables. In this paper, the GHM of each class is calculated based on the grouped KNNs. Given $p, q \geqslant 0$ and a set of $X = \{x_1, x_2, \cdots, x_n\}$, the GHM of X can be defined as

$$GHM^{p,q}(X) = \left(\frac{2}{n(n+1)} \sum_{i=1}^{n} \sum_{j=i}^{n} x_i^{\,p} x_j^{\,q} \right)^{\frac{1}{p+q}} \tag{7}$$

where each x_i is normalized to have a range of values between 0 and 1 using min-max normalization.

Since the unclassified sample has a degree of association to all the classes that are available, a fuzzy membership degree which provides a level of confidence in the classification of the accompanying resultants is calculated. The class with the highest membership value is the predicted class. The fuzzy membership degree can be defined as

$$FM_c(y) = \frac{\sum_{c=1}^{l} u_c \left(\frac{1}{\| y - GHM_c \|^{\frac{2}{m-1}}} \right)}{\sum_{c=1}^{l} \left(\frac{1}{\| y - GHM_c \|^{\frac{2}{m-1}}} \right)} \tag{8}$$

where l is the number of classes, GHM_c is the generalized Heronian mean vector for class c, and u_c is 1 for related class and 0 for other classes.

In this paper, GHM-FKNN algorithm and all experiments were carried out using Python (v3.6.4) and scikit-learn (v0.24.2) package. To avoid overfitting and build the best prediction model, the grid search strategy was applied to select the optimal parameter pair K and (p, q) based on the training dataset.

1.3.2 Evaluation metrics

Four evaluation metrics, namely sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthews correlation coefficient (MCC) are used to evaluate the performance of predictors. These metrics can be formulated as follows.

$$Sn = \frac{TP}{TP + FN} \tag{9}$$

$$Sp = \frac{TN}{TN + FP} \tag{10}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}} \quad (12)$$

where, TP (true positive) is the number of downtrend labels predicted as downtrend labels; FP (false positive) is the number of uptrend labels predicted as downtrend labels; TN (true negative) is the number of uptrend labels predicted as uptrend labels; and FN (false negative) is the number of downtrend labels predicted as uptrend labels.

1.3.3 Model selection and evaluation

10-fold cross-validation test was used to choose the optimal parameters on the training datasets across three stocks. Since each round of cross-validation involves randomly partitioning the original dataset into a training set and a testing set, the random seed value of 1996 was used to obtain repeatable splits when creating folds for cross-validation. 22 pairs of parameters K and (p, q) values were tried as $K \in \{18, 25, 28, 35, 38, 45, 48, 65, 68, 85, 88\}$ and $(p,q) \in \{(1,5), (5,1)\}$, in which the combination of K and (p,q) corresponding to the highest cross-validated Acc value was regarded as the optimal parameter pair. It is worth noting that K is the number of nearest neighbors and (p,q) is the parameters pair for the generalized Heronian mean. The model with the optimal combination of parameters for each stock is then considered the best, i.e., the final prediction model of the proposed predictor GHM-FKNN (Fig. 1). Finally, the prediction models of RF, PRF, KNN, FKNN, and GHM-FKNN are evaluated using independent test datasets for three stocks, respectively.

2 Experimental results and analysis

2.1 Benchmark datasets

Three datasets corresponding to three stocks are used to build the prediction model. These stocks were also used in Ref. [15], namely Apple Inc. (AAPL), Amazon.com Inc. (AMZN) and Netflix Inc. (NFLX). Each dataset consists of two parts: financial data and financial news data, whose time period ranges from 2016-01-01 to 2020-04-01. Before applying to the prediction model, data preprocessing and feature encoding were implemented on three datasets (Section 1). The final dataset for AAPL, AMZN and NFLX consisted of 9 features with a total number of 1064, 1063, and 1066 instances, respectively. All final datasets were then split into 80% training and 20% test sets for prediction model construction and evaluation (Table 1).

Table 1 Stock market datasets used in this experiments

Dataset	Training dataset			Independent test dataset		
	Total number	Positive	Negative	Total number	Positive	Negative
AAPL	870	406	464	194	85	109
AMZN	869	391	478	194	82	112
NFLX	871	425	446	195	95	100

2.2 Model establishment

In order to avoid overfitting and select the optimal parameters for the prediction models, various hyper parameters of the five compared methods (RF, PRF, KNN, FKNN, and GHM-FKNN) were tested to train the models, better adjust the model, and achieve higher accuracy by 10-fold cross-validation test (Table 2). The final prediction model of each compared methods were then constructed by using the corresponding optimal parameter combinations which could achieve the highest cross-validated accuracy value. For the proposed method GHM-FKNN, the impact of 22 combinations of parameters K and (p, q) were investigated on predictive performances of all three datasets (Section 1). For the AAPL dataset, the candidate model with the optimal parameters $K = 45$ and $(p, q) = (1, 5)$ achieved the highest accuracy value of 54.94%, which was considered as the predictive model. For the AMZN dataset, the highest accuracy value of 55.12% was reached by $K = 68$ and $(p, q) = (5, 1)$, which was applied to establish the final prediction model. For the NFLX dataset, it achieved a much lower accuracy value than the other two datasets. The highest accuracy value of 51.79% was obtained by using parameters $K = 88$ and $(p, q) = (1, 5)$.

In recent years, a series of improved Heronian mean (HM) based aggregation operators have been utilized to solve some multi-criteria decision-making problems[19], such as GHM, generalized weighted Heronian mean (GWHM), and improved generalized weighted Heronian mean (IGWHM). To further investigate the effect of HM application on stock trend prediction, GWHM based FKNN (GWHM-FKNN) predictor was implemented and tested using the corresponding optimal parameter combinations of GHM-FKNN (Table 2). Ten different random seeds between 1 and 10 were set to generate ten groups of repeatable weights vector $\boldsymbol{W} = (w_1, w_2, \cdots, w_n)^{\mathrm{T}}$, where $w_i \geqslant 0 (i = (1, 2, \cdots, n))$ and $\sum_{i=1}^{n} w_i = 1$. Given $p, q \geqslant 0$ and a set of $X = \{x_1, x_2, \cdots, x_n\}$, the GWHM of X can be defined as

$$GWHM^{p,q}(X) = \left(\frac{2}{(n+1)} \sum_{i=1}^{n} \sum_{j=i}^{n} (w_i x_i)^p (w_j x_j)^q\right)^{\frac{1}{p+q}} \quad (13)$$

The comparison results showed that GHM-FKNN and GWHM-FKNN achieved relatively similar performance on the training datasets using 10-fold cross-validation test, but the classification accuracy of GWHM-FKNN vary depending upon uncertain weights (Fig. 2). Therefore, in this study, to avoid additional weights determination and complex computation, GHM-FKNN was chosen for final model establishment.

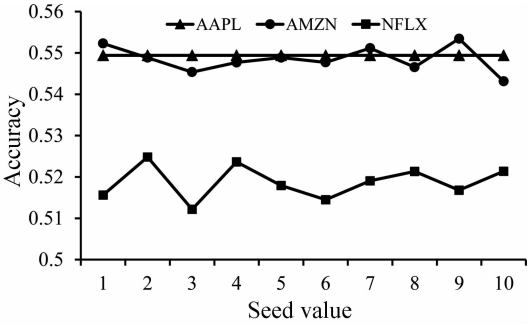

Fig. 2 The classification accuracy of GWHM-FKNN using ten different random seeds

In this study, to explore the relationship between financial news and stock trend, the effect of sentiment score on classification accuracy was investigated (Fig. 3). In the legend of Fig. 3, 'S_used' represents the model using the feature of sentiment score; 'S_unused' represents the model without sentiment score as additional input features. The results showed that the prediction performance of GHM-FKNN can be improved when adding additional feature sentiment scores on the training datasets of AAPL and AMZN. However, at some point, sentiment scores cause performance degradation on the NFLX training datasets. This is to say, the conservatism of the trend varies from stock to stock.

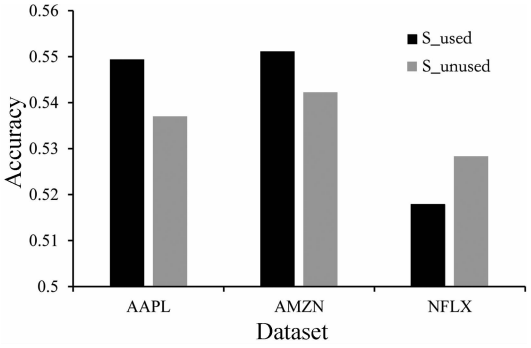

Fig. 3 The effect of sentiment score on classification accuracy.

Table 2 The optimal grid parameters of compared methods on three training datasets

Method	Dataset	Grid parameters	Optimal parameters
RF	AAPL	maxDepth: [2, 3, 4], numTrees: [5, 8, 10, 12], minInstancesPerNode: [1, 2, 3, 4], minInfoGain: [0, 1, 2], featureSubsetStrategy: [" auto", " all", " sqrt", "log2"], impurity: [" entropy", "gini"]	maxDepth:4, numTrees:10, minInstancesPerNode:3, minInfoGain:0, featureSubsetStrategy:" auto", impurity:"entropy"
	AMZN		maxDepth:4, numTrees:5, minInstancesPerNode:2, minInfoGain:0, featureSubsetStrategy:" auto", impurity:"gini"
	NFLX		maxDepth:3, numTrees:12, minInstancesPerNode:2, minInfoGain:0, featureSubsetStrategy:"auto", impurity:"entropy"
PRF	AAPL	*n*_estimators: [5, 8, 10, 12], max_depth: [2, 3, 4]	*n*_estimators:12, max_depth:2
	AMZN		*n*_estimators:8, max_depth:2
	NFLX		*n*_estimators:12, max_depth:4
KNN	AAPL	*n*_neighbors: [18, 25, 28, 35, 38, 45, 48, 65, 68, 85, 88], weights: [" uniform", " distance"], algorithm: [" auto", "ball_tree", "kd_tree", "brute"]	*n*_neighbors: 68, weights = "uniform", algorithm = "auto"
	AMZN		*n*_neighbors: 48, weights = "uniform", algorithm = "auto"
	NFLX		*n*_neighbors: 45, weights = "uniform", algorithm = "auto"
FKNN	AAPL	*n*_neighbors: [18, 25, 28, 35, 38, 45, 48, 65, 68, 85, 88]	*n*_neighbors:48
	AMZN		*n*_neighbors:28
	NFLX		*n*_neighbors:68
GHM-FKNN	AAPL	*K*: [18, 25, 28, 35, 38, 45, 48, 65, 68, 85, 88], [*p*, *q*]: [[1, 5], [5, 1]]	*K*:45, [p, q]:[1, 5]
	AMZN		*K*:68, [p, q]:[5, 1]
	NFLX		*K*:88, [p, q]:[1, 5]

2.3 Comparisons with existing predictors

To further demonstrate the performance of the GHM-FKNN predictor, GHM-FKNN was compared with four existing predictors RF, PRF, KNN, and FKNN. Under 10-fold cross-validation test, the proposed GHM-FKNN method achieved relatively similar performances to RF, PRF, KNN, and FKNN on the training datasets of three stocks

(Table 3). It was demonstrated that Acc, MCC, Sn, and Sp of GHM-FKNN are higher than those of RF, PRF, KNN, and FKNN on the independent test datasets of three stocks or relatively more comparable with them (Table 4). For GHM-FKNN, compared with RF, PRF, KNN, and FKNN, the Acc values of 9.28%, 8.25%, 5.15%, 8.25% improvements were observed on the AAPL dataset, 3.61%, 1.55%, 4.13%, 8.77% improvements were observed on the AMZN dataset, and 0.52%, 11.79%, 9.74%, 12.82% improvements were observed on the NFLX dataset. The above evaluated results clearly illustrated that GHM-FKNN is superior to RF, PRF, KNN, and FKNN.

Table 3 Comparison of five predictors on the training datasets using 10-fold cross-validation test

Dataset	Predictor	*Sn*/%	*Sp*/%	*Acc*/%	*MCC*
AAPL	RF	37.68	85.99	63.45	0.27
	PRF	15.22	86.12	52.87	0.01
	KNN	25.64	83.08	56.32	0.10
	FKNN	45.43	56.91	51.26	0.02
	GHM-FKNN	37.20	70.71	54.94	0.08
AMZN	RF	33.50	86.61	62.72	0.24
	PRF	18.21	84.62	55.00	0.04
	KNN	17.61	85.78	54.76	0.05
	FKNN	38.38	58.73	49.47	−0.03
	GHM-FKNN	21.66	83.26	55.12	0.06
NFLX	RF	NA	NA	65.5	NA
	PRF	40.14	66.35	53.04	0.07
	KNN	50.53	56.77	53.17	0.07
	FKNN	50.80	54.62	52.69	0.05
	GHM-FKNN	48.56	56.30	51.79	0.05

Results excerpted from Ref. [11].

Table 4 Comparison of five predictors on the independent testing datasets

Dataset	Predictor	*Sn*/%	*Sp*/%	*Acc*/%	*MCC*
AAPL	RF	25.88	74.31	53.09	0.01
	PRF	17.65	**82.57**	54.12	0.01
	KNN	32.94	76.15	57.22	0.10
	FKNN	**48.24**	58.72	54.12	0.07
	GHM-FKNN	**48.24**	73.39	**62.37**	**0.22**
AMZN	RF	29.27	73.21	54.64	0.03
	PRF	20.73	**83.04**	56.70	0.05
	KNN	18.29	80.36	54.12	−0.02
	FKNN	**39.02**	57.14	49.48	−0.04
	GHM-FKNN	29.27	79.46	**58.25**	**0.10**
NFLX	RF	NA	NA	63.58	NA
	PRF	35.79	68.00	52.31	0.04
	KNN	49.47	59.00	54.36	0.09
	FKNN	46.32	56.00	51.28	0.02
	GHM-FKNN	**50.53**	**77.00**	**64.10**	**0.29**

Results excerpted from Ref. [11].

3 Conclusion

Stock trend prediction has always been an active and tricky area of the research. In this paper, a new-classifier called GHM-FKNN is proposed for accurate prediction of stock trends using a fuzzy k-nearest neighbor model based on generalized Heronian mean. The comparison results elucidated that GHM-FKNN achieved the best performance with *Acc* of 62.37% for AAPL, 58.25% for AMZN, and 64.10% for NFLX, outperforming four existing predictors RF, PRF, KNN and FKNN on independent test datasets. GHM-FKNN may become a useful tool for stock market analysis. It is important to note that the accuracy of stock trend predictions by GHM-FKNN and other predictors may not be so high due to the chaotic nature of stock prices, insufficient publicly available stock news data and too small benchmark datasets, etc. Fortunately, the growth of financial big data and the development of data mining have brought an opportunity to solve this problem. Given that transformer's attention mechanism was successfully exploited in some previous studies[15, 20], the Transformer-based model may show its values in classification tasks and there are several points in the training process that have further optimization potential. Whether the above problems can be solved perfectly is currently being explored by some other machine learning algorithms and results are expected to be released in the next version of GHM-FKNN in future.

References

[1] TSAI C F, HSIAO Y C. Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches [J]. Decision Support Systems, 2010, 50(1): 258-269.

[2] FAMAE F. Efficient capital markets: a review of theory and empirical work [J]. The Journal of Finance, 1970, 25(2): 383-417.

[3] FAMAE F. Random walks in stock market prices [J]. Financial Analysts Journal, 1965, 21(5): 55-59.

[4] ARÉVALO R, GARCÍA J, GUIJARRO F, et al. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting [J]. Expert Systems with Applications, 2017, 81(9): 177-192.

[5] NGUYEN T H, SHIRAI K, VELCIN J. Sentiment analysis on social media for stock movement prediction [J]. Expert Systems with Applications, 2015, 42(24): 9603-9611.

[6] KALYANI J, BHARATHI H N, JYOTHI R. Stock trend prediction using news sentiment analysis [EB/OL]. (2016-07-07) [2022-09-06]. https://arxiv.org/abs/1607.01958.

[7] SHAH D, ISAH H, ZULKERNINE F. Stock market analysis: a review and taxonomy of prediction techniques [J]. International Journal of Financial Studies, 2019, 7(2): 1-22.

[8] BHURIYA D, KAUSHAL G, SHARMA A, et al. Stock

market prediction using a linear regression[J]. Communication and Aerospace Technology (ICECA), 2017, 2(1): 510-513.
[9] KIM S H, LEE H S, KO H J, et al. Pattern matching trading system based on the dynamic time warping algorithm[J]. Sustainability, 2018, 10(1):1-18.
[10] KHAN W, GHAZANFAR M A, AZAM M A, et al. Stock market prediction using machine learning classifiers and social media, news[J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(1):3433-3456.
[11] KABBANI T, USTA F E. Predicting the stock trend using news sentiment analysis and technical indicators in Spark [EB/OL]. (2022-01-19) [2022-09-06]. https://arxiv.org/abs/2201.12283.
[12] MERVIN L H, TRAPOTSI M A, AFZAL A M, et al. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty[J]. Journal of Cheminformatics, 2021, 62(13):1-17.
[13] WILDER J W. New concepts in technical trading systems [M]. Winston-Salem: Trend Research, 1978.
[14] MUDINAS A, ZHANG D, LEVENE M. Market trend prediction using sentiment analysis: lessons learned and paths forward[EB/OL]. (2019-03-13) [2022-09-06]. https://arxiv.org/pdf/1903.05440.pdf.
[15] LIU H. Leveraging financial news for stock trend prediction with attention-based recurrent neural network[EB/OL]. (2018-11-15) [2022-09-06]. https://arxiv.org/abs/1811.06173.
[16] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1):1-37.
[17] COVER T M, HART P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1):21-27.
[18] WIDYADHANA A, BAGUS C, PUTRA P, et al. A Bonferroni mean based fuzzy K-nearest centroid neighbor classifier[J]. Journal of Computer Science and Information, 2021, 14(1):65-71.
[19] YU D J, WU Y Y. Interval-valued intuitionistic fuzzy Heronian mean operators and their application in multicriteria decision making[J]. African Journal of Business Management, 2012, 6(11):4158-4168.
[20] ZHANG Q Y, QIN C, ZHANG Y F, et al. Transformer-based attention network for stock movement prediction [J]. Expert Systems with Applications, 2022, 202(9): 117239.

WU Zhenfeng, born in 1991. He received his Ph. D degree in 2019 from School of Mathematical Sciences, Nankai University. He is an associate research fellow at Institute of Scientific and Technical Information of China. His research interests include big data and information security, artificial intelligence, and bioinformatics.