ter results in the field of stock price prediction. Stock price prediction, as one of the research hotspots in the field of data mining, is more reliable and accurate based on data-driven methods. Numerical data contains historical trading data such as stock opening and closing prices and volume, and textual data contains stock-related financial news and stock bar comments. The data-driven approach automatically learns the parameters of statistical or machine learning models from numerical and textual data, which effectively reduces the cost of manual forecasting and improves the accuracy and robustness of forecasting results.

This paper first introduces the background knowledge related to stock price forecasting in Section 2. This paper collects and organizes literature related to stock price forecasting, and reviews domestic and international research related to stock price forecasting from two perspectives. Stock price forecasting is divided into statistical analysis methods, traditional machine learning-based methods and deep learning-based methods according to the perspective of different models in Section 3. Stock price forecasting methods based on numerical data and text data according to the source of information are summarized and analyzed in Section 4. The current research challenges and possible future breakthroughs in stock price forecasting are also analyzed in Section 5. Finally, Section 6 concludes the whole paper.

2. Background knowledge

2.1. Problem description

Stock price forecasting is based on accurate statistical data and historical information of the stock market. Based on the history, current situation and laws of the stock market, researchers use reasonable methods to forecast the data of stocks and their trends in the future period. Its essence is financial time series analysis. In stock price forecasting, there usually exist several historical trading data such as opening price, closing price, volume and turnover. The time series of stock prices over a continuous period of time is noted as

$$C_1, C_2, C_3, \ldots, C_t, C_{t+1} \tag{1}$$

where C_t denotes the price of the stock on day t, and C_{t+1} is the forecast price of the stock on the next day. If $C_{t+1} > C_t$, the stock price rises; if $C_{t+1} = C_t$, the stock price remains unchanged; if $C_{t+1} < C_t$, the stock price falls.

2.2. Dataset resources

In terms of resources, stock price prediction datasets can be divided into two types based on information sources: numerical data-based and numerical and text data-based. Among them, numerical data-based datasets contain only numerical data, while numerical and text data-based datasets contain data from both numerical and textual information. Among the datasets based on numerical data such as SSE Composite Index (000001) (Fu and Li, 2012; Huang and Liu, 2019; Liu et al., 2018; Pan and Ding, 2000; Wu and Yang, 2013; Zhang et al., 2012), CSI 300 Index (000300) (Cui et al., 2019; Hu, 2021; Shi and Gao, 2015; Song et al., 2019; Zhang, 2020; Zhou, 2018), S&P 500 index (Chen and Huang, 2021; Ding et al., 2014; Hoseinzade and Haratizadeh, 2019; Schumaker and Chen, 2009; Shen et al., 2018; Wang et al., 2021), time-series data of the CSI 500 Index (000905) (Bao et al., 2020), Hengseng (HSI) index (Shen et al., 2018; Tang and Sun, 2003), and closing price datasets of eight stocks including APA and TSLA (Kan, 2019). The datasets based on numerical and textual data include closing price data and related messages for XOM, DELL, EBAY, IBM, KO (Nguyen and Shirai, 2015), Alibaba stock data and related news information data (Zhang et al., 2018), etc. Table 1 summarizes some representative dataset resources for stock price prediction.

2.3. Stock price forecasting platform

As financial stock markets continue to mature and stock price forecasting research continues to emerge, researchers have now developed a large number of stock price forecasting platforms and tools. Some of the published stock price forecasting platforms are shown in Table 2.

2.4. Technical specifications

Technical indicators are based on statistics derived from stock trading data as an analysis system. The trading data includes stock closing price, turnover rate, volatility and other factors. The main purpose of technical indicators is to analyze the market pattern of stock price fluctuations and to determine the main trends and timing of buying and selling. The commonly used technical indicators are shown in Table 3.

2.5. Evaluation indicators

The so-called evaluation indicators are the criteria for evaluating a model's ability to predict data as well as its high or low generalization ability, and there are six main indicators as follows:

(1) Accuracy is the most commonly used evaluation index, describing the ratio of the number of samples with accurate predictions to the number of all samples. Its expression is shown in Eq. (2):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP and TN are the data judged to be correct, FP and FN are the data judged to be incorrect.

(2) Mean absolute error (MAE) indicates the average of the absolute deviation of each predicted value. Its expression is shown in Eq. (3):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |forecast(i) - actual(i)| \tag{3}$$

where $forecast(i)$ is the predicted value of the stock on a given day, $actual(i)$ is the actual value of the stock on a given day, and n is the sample size of the test data set.

(3) The expression of mean absolute percentage error (MAPE) is shown in Eq. (4):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{forecast(i) - actual(i)}{actual(i)}\right| \tag{4}$$

(4) Mean square error (MSE) is the mean value of the sum of squares of the errors of deviations of the predicted values from the original values. The smaller the value, the more accurate the model prediction is. Its expression is shown in Eq. (5):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} ((forecast(i) - actual(i))^2 \tag{5}$$

(5) Root mean squared error (RMSE) is a deformation of MSE, which is a good measure of how much the data deviate from the true value. Its expression is shown in Eq. (6):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ((forecast(i) - actual(i))^2} \tag{6}$$

(6) Sharpe ratio is used to provide a comprehensive description of the risk and reward of a portfolio. The higher the value, the better the performance of the stock or portfolio. Its expression is shown in Eq. (7):

$$Sharperatio = \frac{E(R_p) - R_f}{\sigma_p} \tag{7}$$

where $E(R_p)$ denotes the payoff rate, R_f denotes the payoff rate of the risk-free investment, and σ_p denotes the standard variance.