Data Science in Finance and Economics

http://www.aimspress.com/journal/dsfe

DSFE, 1(4): 362–392.
DOI: 10.3934/DSFE.2021020
Received: 02 December 2021
Accepted: 20 December 2021
Published: 29 December 2021

Review

A review of data mining methods in financial markets

Haihua Liu[1], **Shan Huang** [1], **Peng Wang**[2] **and Zejun Li**[2*]

[1] Business School of Hunan Institute of Technology, Hengyang 421002, Hunan, China

[2] College of Computer Science and Engineering, Hunan Institute of Technology, Hengyang 421002, Hunan, China

* **Correspondence:** Email: hhliulw@163.com.

Abstract: Financial activities are closely related to human social life. Data mining plays an important role in the analysis and prediction of financial markets, especially in the context of the current era of big data. However, it is not simple to use data mining methods in the process of analyzing financial data, due to the differences in the background of researchers in different disciplines. This review summarizes several commonly used data mining methods in financial data analysis. The purpose is to make it easier for researchers in the financial field to use data mining methods and to expand the application scenarios of it used by researchers in the computer field. This review introduces the principles and steps of decision trees, support vector machines, Bayesian, K-nearest neighbors, k-means, Expectation-maximization algorithm, and ensemble learning, and points out their advantages, disadvantages and applicable scenarios. After introducing the algorithms, it summarizes the use of the algorithm in the process of financial data analysis, hoping that readers can get specific examples of using the algorithm. In this review, the difficulties and countermeasures of using data mining methods are summarized, and the development trend of using data mining methods to analyze financial data is predicted.

Keywords: classification; clustering; data mining; economics; financial markets; forecast; stock markets

JEL Codes: G15, C22

1. Introduction

With the rapid development of information technologies, such as the mobile Internet and the Internet of Things, the total amount of data in human society has been exploding at an alarming rate. IDC is expected 175 Zettabytes of data worldwide by 2025 Patrizio. The huge amount of data and its rapid growth pose huge challenges to traditional database analysis techniques and methods. The traditional data analysis methods that rely heavily on manpower can no longer effectively process and analyze large amounts of data Jagadish et al. (2014); Oussous et al. (2018). The amount of data in various industries is increasing rapidly, which may greatly affect scientific research, industry, commerce, social management and other fields in the future. At present, both academia and industry have extremely high expectations for the value of data Howe et al. (2008); Kum et al. (2011); Shamim et al. (2019). The development of big data technology has been promoted by the government as a national strategy Gamage (2016). With the rapid increase in the amount of data, computing technologies represented by data mining, artificial intelligence, and cloud-edge collaborative computing are gradually becoming the main methods in the field of data processing and analysis Lee (2017); Liu et al. (2021b). In order to improve the discovery of the value of data, researchers have been paying attention greatly on data mining methods. As a new interdisciplinary field, data science is developing rapidly Provost and Fawcett (2013); Iqbal et al. (2020).

The economic operation and social organization of modern society are increasingly inseparable from the financial market, which plays a very important role in the historical process of human development Fields (2017); Daugaard (2019). Financial activities have a profound impact on the current economic development of many countries around the world, which promotes the development of the world economy Lin et al. (2012). There are many factors that affect the financial market Braun (2018). With the development of information technology, more and more financial data can be collected. In the financial market, the success of investors largely depends on the quality of the collected data and the value discovery of these data Nassirtoussi et al. (2014); Goyal and Kumar (2020). Due to its importance to investors and social development, data analysis and research on financial markets have received high attention in the fields of finance, information, and mathematics respectively in the past few decades Yoo et al. (2005). With the rapid increase in the amount of data, traditional methods are becoming more and more inadequate. Therefore, some data mining algorithms have been proposed to provide decision support for investors in different financial markets Carta et al. (2021).

Most of the data in the financial market can be classified as time series data. Due to its inherent dynamic, highly noisy, non-linear and complex chaotic nature, it has always been regarded as a major challenge in the fields of finance, data mining and artificial intelligence Si and Yin (2013); Yang et al. (2020). In the past few decades, decision support systems have been proposed to predict financial time series. These methods can be divided into statistical models and machine learning methods Goyal and Kumar (2020); Wang et al. (2011b). Traditional statistical methods usually assume that the time series studied are generated from a linear process, and try to model the potential time series generation process in order to predict the future value of the series Kumar and Murugan (2013). It is becoming more and more powerless to deal with huge and complex financial data. Therefore, the current data mining methods represented by machine learning are more and more popular among researchers Roychowdhury et al. (2019).

This survey focuses on the application of data mining technology in data analysis in the financial market, and reviews data mining methods that solve the problem of data value discovery in this field. Although a large number of scientific research papers on data mining methods in financial markets have been proposed to solve related problems, there have been few review articles on related content in recent years. The only few review articles often focus on specific financial market applications or each specific data mining algorithm.

The rest of this article is organized as follows. Section 2 discusses some other surveys related to this work. Section 3 describes some basic concepts that represent the background of this research, respectively summarizes the data mining methods to be investigated, and selects the main studies to be reviewed. These main studies are grouped according to the data mining methods. Section 4 discusses the challenges and promising issues in this research area. Section 5 provides a summary of this article.

2. Related work

Data mining with searchable chapters was first proposed in the 1980s and was produced along with knowledge discovery. But the predecessor of data mining can be traced back to earlier stages such as data collection. The term "Knowledge Discovery in Databases" (KDD) first appeared in the Symposium of the 11th International Joint Conference on Artificial Intelligence held in Detroit, USA in 1989, followed by KDD in 1991, 1993 and 1994. Symposium. In 1995, the first International Conference on Knowledge Discovery and Data Mining was held in Canada. Since 1997, KDD has owned a special magazine "Knowledge Discovery and Data Mining". Many foreign research results and papers have been published in this area, and a large number of data mining software has been developed and a large number of related websites have been established. The research of KDD and data mining has become a hot topic in the computer field. In short, need is the mother of invention, because the actual data and information explosion put forward requirements for data mining, so data mining came into being Olson (2006).

The 1960s was the stage of data collection. At this stage, it is limited by data storage capacity, especially when it was still in the stage of disk storage. Therefore, the main solution was the problem of data collection, and more focused on the collection and display of static data. The commercial problems solved were also Based on statistical data based on historical results. The 1980s was the data access stage. The emergence of relational databases and structured query languages has made dynamic data query and presentation possible. People can use data to solve some more focused business problems. At this stage, with the emergence of KDD, data mining has entered Historical stage. The 1990s was the stage of decision-making and support for data warehouses. The rapid development of OLAP and data warehouse technology has made multi-level data backtracking and dynamic processing a phenomenon. People can use data to obtain knowledge and make business decisions. Now is the data mining stage. The great development of computer hardware and the emergence of some advanced data warehouses and data algorithms have made it possible to process and analyze massive amounts of data. Data mining can help solve some predictive problems Han et al. (2007); Liao et al. (2012).

We start our investigation by examining review articles on the application of data mining methods to financial market issues, and there are not many review articles covering this topic. Most of them were concentrated in the stock market, but some of them described other financial issues such as exchange rate forecasts, financial crisis analysis, credit scoring, and interest rate forecasts. Most of the published surveys found in the literature were dedicated to the application of certain data mining methods in the financial market, such as decision trees, ensemble learning, artificial neural networks, support vector machines, text mining, and bionic computing.

Some review works compared most relevant forecasting techniques. Atsalakis and Valavanis Atsalakis and Valavanis (2009) reviewed more than 100 scientific literature on data mining methods to solve stock market forecasting problems from the four characteristics of predictor variables, modeling techniques, benchmarks, and performance metrics. They summarized neural networks and fuzzy networks by describing the types of transfer functions, types of membership functions, network architecture and training methods. Soni Soni (2011) investigated the application of artificial neural networks (ANN) in stock market forecasting. Before enumerating some major researches that use ANN methods to solve stock market forecasting problems, this paper provides a brief history of ANN and the basic concepts of the stock market. Nassirtoussi et al. Nassirtoussi et al. (2014) reviewed text mining for market forecasting, they classify and summarize the articles according to the nature of the text, market category, pre-processing procedures, and type of modeling technology. Nardo et al. Nardo et al. (2015) searched the literature for works linking changes in stock returns. They demonstrated the limitations of these works and proposed some modeling techniques for future research. Cavalcante et al. Cavalcante et al. (2016) not only analyzed forecasting algorithms, but also summarized various algorithms related to other problems in the financial market. They proposed feature selection, clustering, segmentation and outlier detection to deal with these problems, and also proposed an architecture for autonomous transactions in the real market. This article reviews the most important research published from 2009 to 2015, which involves the application of several data mining and computational intelligence methods in a variety of financial applications, including the preprocessing and clustering of financial data, and predicting the future market Trends, mining financial text information and other technologies. Xing et al. Xing et al. (2017) selected the most important documents related to financial forecasting based on natural language. They listed the types of financial text used as predictive input and their processing methods, and described the algorithms involved in modeling and implementation details. Bustos and QuimbayaBustos and Pomares-Quimbaya (2020) aimed to update and summarize the forecasting techniques used in the stock market, including classification, characterization and comparison. The focus of this review is the study of the forecast of stock market trends from 2014 to 2018. In addition, it analyzes surveys and other reviews of recent studies published in the same time frame and in the same database. Although these reviews have intersections with our research period and topics, their scope is limited because they tend to focus on only one family of data mining algorithms. Our review is broader. Our review also covers the application of major data mining algorithms in the financial market, and describes and introduces data mining algorithms. Compared to these comments, this review provides an up-to-date discussion on the topic. Given the large number of new works published every year, these surveys will quickly become obsolete.

3. Data mining method review in financial markets

A large amount of data is generated in the financial market every day, and these data contain valuable information. If the value of these data can be discovered through data mining methods, it is of great significance to the healthy operation of the financial market and guiding investors to make investment choices. In this part, we summarize the principles and steps of eight commonly used data mining methods (decision trees, support vector machines, Bayesian method, K-nearest neighbors, K-means, Expectation-maximization algorithm, ensemble learning) Wu et al. (2007); Han et al. (2000), and respectively comprehensively summarize the application of these methods in financial market data analysis. These data mining algorithms are all proposed to deal with data analysis problems. They have their own advantages and disadvantages and are suitable for different scenarios. They are summarized in Table 1.

Table 1. Summary of advantages and disadvantages of the algorithm and applicable scenarios.

Algorithms	Advantages	Disadvantages	Applicable scenarios
C4.5	The computational complexity is not high, the output result is easy to understand, and it is not sensitive to the lack of intermediate values;	may cause over-matching problems;	Numerical and nominal
SVM	Low generalization error rate, low computational overhead, and easy interpretation of the results;	Sensitive to parameter adjustment and selection of kernel function;	Numerical and nominal
Bayesian	It is still valid when there is less data, and it can handle multi-category problems;	More sensitive to the preparation of input data;	Nominal
KNN	High accuracy, insensitive to outliers, no data input assumptions;	High computational complexity and high space complexity;	Numerical
K-means	Easy to implement;	May converge to a local minimum, slower convergence on large-scale data sets;	Numerical
EM	No need to optimize the step length, fast training speed;	The structure of the model cannot be faulted, and the global optimum cannot be found;	Numerical
AdaBoost	Low generalization error rate, easy to code, can be applied to most classifiers, no parameter adjustment;	Sensitive to outliers;	Numerical and nominal

3.1. Decision trees

3.1.1. Introduction to decision trees

Decision trees have always been very popular in the field of data mining due to their simplicity, ease of use and powerful functions. They can be used for classification or regression problems. The decision tree is essentially abstracted as a series of if-then statements, the output of which is a series of simple and practical rules. The decision tree model classifies data instances based on attributes. Each non-leaf node represents a test on a feature attribute, and each branch represents the output of this feature attribute in a certain value range, and each leaf node stores one category. The process of using a decision tree to make a decision is to start from the root node, test the corresponding feature attributes in the items to be classified, and select the output branch according to its value until the leaf node is reached, and the category stored in the leaf node is used as the decision result Quinlan (1986); Kotsiantis (2011).

The construction process of decision tree is generally divided into three parts, namely feature selection, decision tree production and decision tree tailoring. The goal of building a decision tree is to learn a decision tree model based on a given training data set. When constructing a decision tree, the regularized maximum likelihood function is usually used as the loss function for model training, and the learning goal is to minimize the loss function. The algorithm for constructing a decision tree usually selects the optimal feature through recursion, and partitions the training data according to the feature. Decision tree models often have better classification capabilities for training data, but in practice we are more concerned about the classification capabilities for unknown data. Therefore, after the decision tree is built, it usually needs to be pruned, so that the decision tree has better generalization ability HSSINA et al. (2014).

Feature selection is to select features with strong classification ability. The classification ability is generally characterized by information gain, information gain ratio or Gini index. The criterion for selecting features is to find the locally optimal feature as a judgment for segmentation, which depends on the degree of order of the categories in the node data set after segmentation. The more ordered the classified data after the division, the more appropriate the division rules. Feature selection means to select a feature from a number of features as the criterion for splitting the current node. There are different quantitative evaluation methods for the selected features, thereby deriving different decision trees. The goal of attribute selection is that after the data set is divided using a certain attribute, the order of each data subset is higher than the order of the data set before the division García et al. (2009).

According to the selected feature evaluation criteria, child nodes are generated recursively from top to bottom, and the decision tree stops growing until the data set is indivisible. This process is actually continuously dividing the data set into subsets with a higher degree of order and less uncertainty using features that meet the partitioning criteria. For each division of the current data set, it is hoped that the order of each subset after division according to a certain feature will be higher and the uncertainty will be smaller.

The model generated by the decision tree is very accurate in predicting the training data, but it is very poor in classifying unknown data. In fact, the condition for stopping the construction of the decision tree is until there are no features to choose from or the information gain is less than a preset small threshold, which leads to the construction of the decision tree model is too complicated. The model is too dependent on the training data, which makes it perform poorly on the test data set, resulting in over-fitting. Over-fitting is often because the decision tree is too complicated. The solution is to control the complexity of the model and simplify the model, which is called pruning technology. The purpose of decision tree pruning is to improve generalization ability through pruning. The idea is to find a balance between the prediction error of the decision tree on the training data and the model complexity, thereby alleviating over-fitting.

The most typical models in the decision tree algorithm include ID3 (selecting features through information gain), C4.5 (through information gain ratio feature selection), CART (feature selection through Gini index) Chen et al. (2014a); Salzberg (1994); Jaworski et al. (2014). Below we take C4.5 as a representative to introduce the algorithm process of the decision tree. The C4.5 algorithm is described as follows Algorithm Alg$_C$4.5.

Algorithm 1 C4.5 Algorithm

```
Input:
    Training set: $D$;
    Attribute set: $A$;
    Threshold of attribute information gain ratio: $\varepsilon > 0$.
Output:
    Decision tree: $T$
    Steps:
1: If all samples in $D$ belong to the same class $c_k$, then $T$ is a single-node tree, and $c_k$ is used as the class label of the node, and $T$ is returned.
2: If $A = \phi$, then $T$ is a single-node tree, take the class $c_k$ with the largest number of samples in $D$ as the class label of the node, and return $T$.
3: Otherwise, $g_R(D, A_i)$ is calculated, where $A_i \in A$ is each attribute in the attribute set, and the feature $A_g$ with the largest information gain ratio is selected.
4: Judging the information gain ratio of $A_g$ is as follows:
        If $g_R(D, A_i) < \varepsilon$, it is a single-node tree, take the class $c_k$ with the largest number of samples in $D$ as the class label of the node, and return $T$;
        If $g_R(D, A_i) >= \varepsilon$, for each possible value of the $A_g$ attribute, $D$ is divided into several non-empty subsets $D_i$ according to $A_g = a_i$, and the class with the largest number of samples in $D_i$ is used as the label class, and the child node is constructed. The child node and its The child nodes form a tree $T$, and $T$ is returned.
5: For the $i$th child node, take $D_i$ as the training set and $A - \{A_g\}$ as the attribute set, call the previous steps recursively, get the subtree $T_i$, and return $T_i$.
```

3.1.2. Application of decision tree in financial market

Decision trees have been widely used in financial markets for their simplicity and ease of use. The author used decision trees obtained through the application of artificial intelligence strategies to classify the converted data, analyzes and evaluates different decision trees, shows accuracy and emphasizes the relevance of capital gains Mir-Juli et al. (2010). Based on the financial statement detection information and detection rules extracted from the C4.5 decision tree algorithm, Tang et al. developed a knowledge-based financial statement fraud detection system to detect fraudulent activities in financial statements and discover hidden knowledge Tang et al. (2018b). Chen et al. used the decision tree classification method and logistic regression technology to implement the financial distress prediction model, and concluded that the closer to the actual occurrence of financial distress, the higher the accuracy of the decision tree classification method, and the correct rate of the two quarters before the occurrence of financial distress Reached 97.01% Chen (2011). By combining text mining technology, feature selection and decision tree algorithms, Nasseri et al. proposed a new intelligent transaction support system based on sentiment prediction to analyze and extract semantic terms expressing specific emotions, and built assumptions based on predetermined investment on this basis. The trading strategy was used to evaluate the profitability of the term trading decision extracted from the decision tree model Nasseri et al. (2015). Zhao et al. used decision tree algorithms to mine and explore the financial business related data of sports industry enterprises, and conduct risk assessment through related indicators. And by using the model to calculate the probability of risk occurrence and analyze the degree of damage Zhao (2021). Malliaris et al. used the decision tree method to study the behavior of gold prices using traditional financial variables such as stock volatility, oil prices, and the euro Malliaris and Malliaris (2015). Based on the core of financial indicators and combined with decision tree technology, Cheng et al. established a hybrid classification model and predictable rules that affect the rise and fall of stock prices Cheng (2014). For the situation when decision makers with heterogeneous preferences facing costly information acquisition and analysis, Luo et al. used decision trees to model the formation of herd behavior in the P2P lending market Luo and Lin (2011). Incorporating heterogeneous information from various sources such as social media, world market performance, financial news, and historical data, Agrawal et al. proposed a new technique to increase the mixing degree of a single decision tree that exists in the forest Agrawal and Adane (2021). Deng et al. proposed a combination of Gradient Boosting Decision Tree and Differential Evolution, using relevant indicator data to identify insider trading activities Deng et al. (2019).

Jan et al. constructed four types of decision trees to establish a rigorous and effective corporate financial statement fraud detection model for the sustainable development of enterprises and financial markets long Jan (2018). Rosati et al. proposed a closed-loop data mining method based on a decision tree model to analyze the results of financial transaction data Rosati et al. (2020). Yeo et al. proposed a decision tree analysis model composed of technology, organization, and environmental frameworks. They found that financial factors are more predictive of service industry performance than information and communication technologies Yeo and Grant (2018). Tsai et al. used the decision tree model to generate useful decision rules to allow investors and creditors to make effective earnings management forecasts Tsai and Chiou (2009). Ohana et al. used the gradient boosting decision tree method to predict a substantial price drop in the Standard Poor's 500 index from a set of 150 technical, fundamental and macroeconomic characteristics Ohana et al. (2021). Kim et al. used majority voting integration methods and decision trees, as well as experimental data from American restaurants from 1980 to 2017, to develop a more accurate and stable business failure prediction model Kim and Upneja (2021). Sun et al. proposed a new decision tree ensemble model based on a small number of oversampling techniques and bagging ensemble learning algorithm with differential sampling rate, which is used for unbalanced enterprise credit evaluation Sun et al. (2018a). Farid et al. tried to use the decision tree model to predict stock prices in emerging markets through analysis using WEKA software Farid et al. (2021). He et al. introduced the basic methods of decision trees and boosted trees, and showed how these classification methods can be applied to the detection of abnormal transactions He et al. (2020). Chen et al. used two decision tree algorithms, CART and C5.0, to construct an effective and innovative two-stage continuous operation prediction model to predict the continuous operation of the enterprise and the development of the capital market Chen (2019). Based on the organization's commitment to logistics financial services and public social media sharing a lot of sentiment related to sales prices and products, Awan et al. used MLlib models such as random forests and decision trees to study the stocks of 10 top companies, and their data includes historical stock prices Awan et al. (2021). It is recommended to analyze the decision tree created by genetic programming, the goal of which is to extract and collect different rules for classifying positive cases to predict future opportunities in the financial stock market Garciaalmanza and Tsang (2006).

3.2. Support vector machines

Support vector machine (SVM) is a new and very potential classification technology proposed by the AT&T Bell laboratory research group led by Vapnik in 1995. At the beginning, it was mainly proposed for the binary classification problem, and successfully applied the sub-solution function regression and the first-class classification problem, and extended it to the actual multi-value classification problem in a large number of applications. SVM is a supervised learning model related to related learning algorithms. Since its birth, SVM has swept the field of machine learning due to its good classification performance, and has been highly valued by researchers Cortes and Vapnik (1995).

3.2.1. Introduction to support vector machines

SVM is a two-classification model. Its basic model is a linear classifier with the largest interval defined in the feature space; the improved SVM also includes kernel techniques, which enable it to handle nonlinearities. The learning strategy of SVM is to maximize the interval, which can be formalized as a problem of solving convex quadratic programming, which is also equivalent to the problem of minimizing the regularized hinge loss function. The learning algorithm of SVM is essentially an optimization algorithm for solving convex quadratic programming Raudys (2000).

The SVM method is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. According to the limited sample information, it seeks the best compromise between the complexity of the model and the learning ability in order to obtain the best promotion ability. SVM can be used for classification or regression. Given a set of training samples with marked as belonging to two categories, the SVM training algorithm builds a model and assigns new instances to one category or other categories to make it a non-probabilistic binary linear classification.

In addition to linear classification, support vector machines can also use kernel techniques, whose input is implicitly mapped into a high-dimensional feature space to effectively perform nonlinear classification. When SVM deals with nonlinear problems, it essentially establishes a hyperplane with maximum separation in a high-dimensional space. Two hyperplanes parallel to each other are built on both sides of the hyperplane that separates the data. Establishing a separating hyperplane with a proper direction maximizes the distance between two hyperplanes parallel to it. The assumption is that the greater the distance or gap between parallel hyperplanes, the smaller the total error of the classifier Smola and Schölkopf (2004). The classification diagram of the SVM hyperplane is as follows Figure 1:

SVM is essentially a non-linear method. In the case of a small sample size, it is easy to learn the non-linear relationship between data and features. Therefore, its advantage is that it can solve non-linear problems and avoid the problem of selecting local minimums. It can solve high-dimensional problems

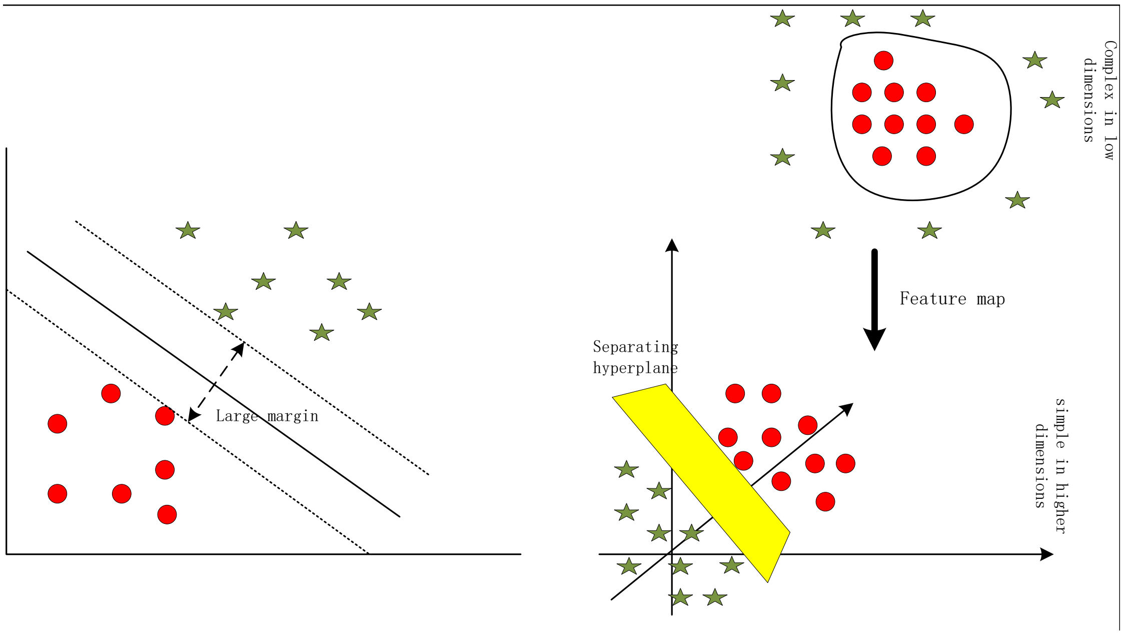

Figure 1. The classification diagram of the large margin and the SVM hyperplane.

and improve generalization performance. At the same time, SVM is very sensitive to lack of data, and depends on the choice of kernel function for non-linear problems. There is no general solution and the computational complexity is high Burges (1998). The SVM is described as follows Algorithm $\mathrm{Alg}_s vm$:

3.2.2. Application of support vector machines in financial market

Ahn et al. extended the classification method used to construct early warning systems for potential financial crises to traditional crises, which proved that SVM is an effective classifier Ahn et al. (2011). Loukeris et al. designed an SVM hybrid evaluation model based on the combination of support vector machines and genetic algorithms. The accounting statements obtained by them provide investors with excellent portfolio choices Loukeris et al. (2013). Kewat et al. applied SVM to predict stock price indices. They tested the feasibility of applying SVM to financial forecasting by comparing SVM with back-propagation neural networks and case-based reasoning Kewat et al. (2017). Liu et al. used the latest intelligent search algorithm Grey Wolf optimizer, combined with a support vector regression machine to design a predictive model, which can obtain the best investment portfolio well Liu et al. (2021a). Based on nuclear regression and support vector machine training, Chen et al. empirically tested the use of machine learning models to hedge the price risk associated with holding energy financial products Chen et al. (2019). Wang et al. proposed a two-step kernel learning method based on support vector regression for financial time series forecasting Wang and Zhu (2008). Guo et al. constructed a predictive model to predict stock market behavior with the help of local preserving projection, particle swarm optimization and support vector machines Zhiqiang et al. (2012). By comparing SVM with multi-layer back propagation neural network and regularized radial basis function neural network to test the feasibility of applying SVM in financial forecasting, Cao et al. discussed the use of SVM in financial time series forecasting Cao and Tay (2003).

Algorithm 2 SVM Algorithm

```
Input:
    Training set: $T = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)\}$,
    Penalty function: $C > 0$.
Output:
    Classification decision function:$f(\mathbf{x})$
    Steps:
1: Choose an appropriate kernel function $K(\mathbf{x}, \mathbf{z})$ to solve the constrained optimization problem:
       $\min_{\alpha} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{m} \alpha_i$
       $s.t. \sum_{i=1}^{m} \alpha_i y_i = 0$                                              (1)
       $C >= \alpha_i >= 0, i = 1, 2, ..., m$
   Find the optimal solution $\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_m^*)^T$.
2: Calculate:
       $\boldsymbol{w^*} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{\alpha_i^* y_i x_i}$                    (2)
   At the same time, select a suitable component $C > \alpha_j^* > 0$ of $\alpha^*$ , and calculate:
       $b^* = y_j - \sum_{i=1}^{m} \alpha_i^* y_i K(x_i, x_j)$                             (3)
3: Construct the classification decision function $f(\boldsymbol{x})$:
       $f(\boldsymbol{x}) = sign(\sum_{i=1}^{m} \alpha_i^* y_i K(x_i, x) + b^*)$                (4)
```

Through logistic regression, support vector machines and random forests, network indicators are used as input for determining strategies to check the effectiveness and usefulness of network indicators, Lee et al. proposed financial network indicators that can be applied to global stock market investment strategies Lee et al. (2019). Schumaker et al. used SVM tailored specifically for discrete numerical forecasts and models containing different stock-specific variables to estimate discrete stock prices 20 minutes after the news article was published Schumaker and Chen (2009). In order to improve the accuracy of short-term financial time series prediction of Markov switched multifractal model (MSM), Wang et al. proposed a SVM-based MSM, which uses the MSM model to predict volatility and uses SVM to innovation for modeling Wang et al. (2011a). By modeling the jumping fluctuations of high-frequency data, Yang et al. designed a prediction model based on SVM to predict short-term fluctuations of high-frequency data Yang et al. (2020). The key to the research of Barboza et al. is to use machine learning technology including SVM to significantly improve the accuracy of prediction, and apply it to the data of North American companies from 1985 to 2013, integrate information from the Solomon Center database and Compustat, and analyze annual observations of more than 10,000 companies Barboza et al. (2017). Taking into account the imbalance between financial distressed enterprises and normal enterprises, Huang et al. further proposed a feature-weighted SVM based on biorthogonal wavelet hybrid kernel for financial distress prediction Huang et al. (2014). Carpinteiro et al. introduced the research of three predictive models-multilayer perceptron, support vector machine and hierarchical model, and trained and evaluated them based on the time series of Brazilian stock market funds Carpinteiro et al. (2011). Kercheval et al. proposed a machine learning framework based on SVM to capture the dynamics of high-frequency limit order books in the financial stock market and automatically predict indicators in real time Kercheval and Zhang (2015). By constructing an integrated model of SVM, Sun et al. focused on how to effectively construct a dynamic financial distress prediction model based on category imbalanced data streams Sun et al. (2020).

3.3. Bayesian method

The Bayesian method is a general term for classification algorithms, which is a classification algorithm developed based on Bayes' theorem. The classification principle of the Bayesian method is generally to analyze the prior probability of a single data, use the Bayesian formula to calculate its posterior probability, and select the category with the largest posterior probability as the category of the data Bishop (2006); Friedman et al. (1997). Approximate solution to Bayes' theorem can provide an effective way for data mining algorithm design. In order to avoid the problem of sample sparseness and combinatorial explosion in solving Bayes' theorem, the naive Bayes classification algorithm introduces the assumption of feature condition independence. Although this assumption is difficult to hold in real applications, Naive Bayes can achieve quite good performance in many situations Ng and Jordan (2002).

3.3.1. Introduction to Bayesian method

The most common Bayesian method is the naive Bayes classifier. In order to avoid the difficulty of directly estimating the joint probability of all attributes from a limited training sample, the naive Bayes classifier adopts the attribute independence hypothesis, that is, each attribute is independent of each other Li et al. (2018). The algorithm is described as follows Algorithm Alg$_{b}ys$:

Algorithm 3 Naive Bayes Algorithm

```
Input:
    Training set: $T = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)\}$, $\mathbf{x_i} = (x_i^1, x_i^2, ..., x_i^n)^T$, $x_i^j$ is the $j$th attribute of the $i$th sample, where $x_i^j \in \{a_{j1}, a_{j2}, ..., a_{js_j}\}$, $a_{jl}$ is the $l$th value that the $j$th attribute may take, $j = 1, 2, ..., n$, $l = 1, 2, ..., s_j$, $y_i \in \{c_1, c_2, ..., c_k\}$.
Output:
    Classification of sample $\mathbf{x}$
    Steps:
1: Calculate the estimated value of the prior probability and the estimated value of the conditional probability:
       $P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1, 2, ..., K$    (5)
       $P(X^j = a_{jl}/Y = c_k) = \frac{\sum_{i=1}^{N} I(x_i^j = a_{jl}, y_i = c_k)}{\sum_{i=1}^{N} I(y_i = c_k)}$    (6)
2: For a given sample $\mathbf{x} = (x^1, x^2, ..., x^n)^T$, calculate:
       $P(Y = c_k) \prod_{j=1}^{n} P(X^j = x^j/Y = c_k)$    (7)
3: Calculate and return the classification $y$ of sample $\mathbf{x}$:
       $y = arg \max_{c_k} P(Y = c_k) \prod_{j=1}^{n} P(X^j = x^j/Y = c_k)$    (8)
```

Compared with other classifiers, Bayesian classifier has stable classification efficiency, that is, it can adapt to discrete attributes and can also be used for continuous attributes. When the attributes are independent of each other, the classification effect is very good, and it also has the characteristics of being less sensitive to missing and noisy data Bielza and Larrañaga (2014).

3.3.2. Application of Bayesian method in financial market

Luintel et al. used a new data set covering 69 countries from 1989 to 2011 in the Bayesian framework to re-examine the issue of whether the financial structure is related to economic growth Luintel et al. (2016). Yan et al. used financial research reports training an LDA model based on the Bayesian framework to predict the most relevant financial news among the 24 primary industries in the Chinese market Yan and Bai (2016). Based on Lending Club's real-world P2P lending data set, which contains a large amount of data collected from 2007 to 2016, Fields et al. used Bayesian learning schemes to analyze how to supervised classification models dealing with category imbalances and to affect credit prediction rates Ferreira et al. (2017). Zhang et al. proposed a new non-linear classification method based on the Bayesian "sum of trees" model, this method extends the Bayesian additive regression tree method to the classification context and is applied to predict whether the company will be insolvent Zhang and Härdle (2010). Based on the financial data of 20 publishing and media companies in 2011, Zhao et al. conducted a cluster analysis of profitability, solvency, growth and capital expansion capabilities, and used Bayesian discriminant analysis to quantitatively test the classification to expect the public investment in stocks become more rational and scientific Zhao et al. (2014). Talebi et al. proposed a new classification method based on the Bayesian method to identify the upward, downward, and sideways trends of exchange rates in the foreign exchange market Talebi et al. (2014). Kanhere et al. applied Bayesian classification technology to abnormal activity risk classification for the purpose of automating the abnormal activity detection and report generation process Kanhere and Khanuja (2015). Zhuang et al. used Bayesian Networks (BN) to deal with the problem of uncertain experience attributes in customer relationships to evaluate customer information scores Zhuang et al. (2015). Pan et al. proposed a new method of learning from data using a rule-based fuzzy inference system, in which Bayesian inference and Markov chain Monte Carlo techniques were used to estimate model parameters. It also shows the applicability of this method in the financial services industry using synthetic data sets for regression and classification tasks Pan and Bester (2018).

Kirkos et al. investigated the usefulness of Bayesian Belief Network in identifying false financial statements, and used it to explore the effectiveness of data mining classification techniques in detecting companies with fraudulent financial statements KIRKOS et al. (2007). Ma et al. used the Naive Bayes algorithm to evaluate data and classify the degree of confidence in the real estate market by this metric to study new trends in finance and economy Ma et al. (2017). Peng et al. constructed a performance score to measure the performance of the classification algorithm, and introduced a Bayesian network-based multi-criteria decision-making method to provide the final ranking of the classifier, the purpose of which is to evaluate financial risks Peng et al. (2011). Ryans et al. used Naive Bayesian classification to identify comment letters related to future restatements and write-downs to examine the impact of comment letters on future financial report results and earnings credibility Ryans (2020). Vaghela et al. proposed the MR-MNBC method based on MaxRel feature selection as the preprocessing task of the multi-relational naive Bayes classifier, and analyzed the modified algorithm on the PKDD financial data set Vaghela et al. (2014). Kim et al. developed the Bayesian network as a predictive tool for detecting and classifying misstatements based on whether there is fraudulent intent, and evaluated the research characteristics of detecting fraudulent intent and major misstatement Kim et al. (2016). Guo et al. proposed a new multi-level adaptive classifier ensemble model based on statistical techniques and Bayesian methods to improve the accuracy of credit ratings and improve risk control and profitability Guo et al. (2019).

3.4. K-nearest neighbors

The KNN classification algorithm is a theoretically mature method, and it is also one of the simplest and best-understood algorithms of all data mining algorithms. KNN is a basic classification and regression method. Its input is the attribute vector of the instance, and the K nearest data is selected for classification and discrimination by calculating the distance between the instance and the attribute value of the training data. KNN does not have an explicit learning process but directly makes predictions, so it is also called a lazy learning classifier. It actually uses the training data set to divide the attribute vector space and classify it Gou et al. (2019).

3.4.1. Introduction to K-nearest neighbors

The KNN algorithm includes K value selection, distance measurement and classification decision rules. The K value market value selects the nearest K neighboring data for distinguishing categories. The choice of k value will have a significant impact on the results of the algorithm. A small k value means that only training examples close to the input instance will have an effect on the prediction result, but overfitting is prone to occur; if the k value is large, the advantage is that it can reduce the estimation error of learning, but the disadvantage is that learning the approximation error increases. At this time, the training instance farther from the input instance will also have an effect on the prediction, making the prediction wrong. In application, the k value generally selects a smaller value, and the cross-validation method is usually used to select the optimal k value. In essence, it is to compare the average cross-validation error rate of different k values, and select the k value with the smallest error rate.

The KNN algorithm requires that all feature attributes of the data can be quantified comparatively. If there are non-numerical attributes in the data features, methods must be used to quantify them as numerical types. Data has multiple attributes, and each attribute has its own domain. Different spans of attribute values will have different effects on distance calculation. A large domain will have a greater impact than an attribute with a small domain, so the attributes must be unified and normalized processing. The distance between two data points in the feature space actually reflects the similarity of the two data. The feature space of KNN is generally an n-dimensional real number vector space R^n. When using the KNN method, Euclidean distance is generally used to measure the similarity of two data, of course, the general L_p distance can also be used Muja and Lowe (2014).

The choice of classification decision usually adopts majority voting, or weighted voting based on the distance. The closer the distance, the greater the weight of the sample. The description of the KNN classification algorithm is as follows Algorithm 4:

The characteristic of the KNN classification algorithm is that there is no explicit training process. It only saves the training data during the training phase, and the training time overhead is zero, until the test data is received before processing Samworth (2012).

3.4.2. Application of K-nearest neighbors in financial market

Zemke et al. merged the improved KNN method into a simple stock exchange trading system to achieve returns higher than exponential growth Zemke (1999). Zhang et al. used the False Nearest Neighbors method to obtain the embedding dimension m in the KNN algorithm, and applied the modified method to the daily adjustment opening value of Tsingtao Brewery Co., Ltd Zhang and Li (2010). Sun et al. developed a new prediction model based on four KNN classifiers, which enhanced the accuracy of

Algorithm 4 K-nearest neighbors Algorithm

```
Input:
    Training set: $T = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)\}$, $\mathbf{x_i} = (x_i^1, x_i^2, ..., x_i^n)^T$, $x_i^j$ is the $j$th attribute of the $i$th sample, where $x_i^j \in \{a_{j1}, a_{j2}, ..., a_{js_j}\}$, $a_{jl}$ is the $l$th value that the $j$th attribute may take, $j = 1, 2, ..., n$, $l = 1, 2, ..., s_j$, $y_i \in \{c_1, c_2, ..., c_k\}$.
Output:
    Classification of sample $\boldsymbol{x}$
    Steps:
    According to the given distance metric, find the k nearest neighbors to $x$ in $T$. Define the neighborhood of $x$ that covers these $k$ points and denote it as $N_k(\boldsymbol{x})$.
    From $N_k(\boldsymbol{x})$, determine the category $y$ of $\boldsymbol{x}$ according to the classification decision rule:
        $y = arg \max_{c_j} \sum_{\boldsymbol{x_i} \in N_k(\boldsymbol{x})} I(y_i = c_j), i = 1, 2, ..., m; j = 1, 2, ..., k \qquad (9)$
    Among them, $I$ is the indicator function: $I(ture) = 1, I(false) = 0$. In the above formula, only the sample points in $\boldsymbol{x_i} \in N_k(\boldsymbol{x})$ are considered for $y_i, i = 1, 2, ..., m$.
```

the denoising process to predict stock market trends Sun et al. (2017). Lin et al. proposed an improved modeling program to improve the function of the KNN method in financial time series forecasting to help investors obtain a reasonable investment portfolio Lin et al. (2021). Nie et al. researched and applied the KNN network based on Frobenius distance to analyze the time-varying characteristics of the correlation matrix, especially during periods of drastic changes, and applied it to the analysis of US stock market data Nie (2020). Taking into account the recent trends and the limitations of previous studies, Aijing et al. proposed a new method called empirical mode decomposition combined with KNN to predict stock indexes LIN et al. (2012). Chen et al. proposed a basic hybrid framework of feature-weighted support vector machines and feature-weighted KNN to effectively predict stock market indexes and obtain more profits and returns with a lower risk rate through effective trading strategies Chen and Hao (2017). In financial distress research, in order to obtain better classification accuracy and maintain data integrity, Cheng et al. proposed a purity-based k-nearest neighbor algorithm to improve the performance of missing value interpolation Cheng et al. (2019). Yu et al. built a violations detection model based on the KNN method to detect violations of listed companies to avoid investment risks Yu et al. (2013).

Chen et al. proposed an effective non-parametric classifier for bankruptcy prediction using an adaptive fuzzy KNN method Chen et al. (2011). Li et al. introduced the KNN principle on each feature and established a new feature-based similarity measurement mechanism to study financial distress prediction including bankruptcy prediction Li et al. (2009). Zhang et al. proposed a new method that combines integrated empirical mode decomposition and multi-dimensional KNN model to predict the closing price and high price of stocks at the same time, which has high prediction accuracy for short-term forecasts Zhang et al. (2017). Based on the fact that the parameters used to generate training patterns are heuristically determined by automatic mutual information and false nearest neighbor methods, Qian et al. proposed an inductive machine learning classifier based on KNN to predict the stock market Qian and Rasheed (2006). Huang et al. proposed the use of bi-clustering mining to discover effective technical transaction patterns for improving the KNN method applied to the classification of trading days during the test period Huang et al. (2015). Tang et al. proposed a new computational intelligent model for predicting time series based on KNN, and a complex prediction model for stock market indexes that integrates all industry index predictions with it as the core Tang et al. (2018a). Li et al. proposed a hybrid decision model that uses case-based reasoning to enhance genetic algorithms and fuzzy KNN methods to predict financial activity rates Li and Ho (2009). Based on the newly designed fuzzy prediction trading system and the non-parametric system of KNN technology, Naranjo et al. proposed a novel fuzzy recommendation system for stock market investors, which has the same all-or-nothing investment strategy and risk control Naranjo and Santos (2019).

3.5. k-means

The K-Means algorithm is an unsupervised clustering algorithm. It is relatively simple to implement and has a good clustering effect, so it is widely used.

3.5.1. Introduction to k-means

The K-Means algorithm is an unsupervised algorithm that focuses on similarity. It uses distance as a measure of the similarity between data objects. That is, the smaller the distance between data objects, the higher their similarity, and the more likely they are in the same cluster. It can find k different clusters, and the center of each cluster is calculated by using the mean of the values contained in the clustersCelebi et al. (2013). The algorithm flow is as follows algorithm 5: The principle of the K-means algorithm is relatively simple, the implementation is also very easy, and the convergence speed is fast. The clustering effect is better, and the interpretability of the algorithm is relatively strong. In addition, the main parameter that needs to be adjusted is only the number of clusters k. In addition to the advantages, the disadvantages of K-means are also obvious. The selection of the k value is not easy to grasp. If the data of each hidden category is unbalanced, or the variance of each hidden category is different, the clustering effect will be not good. The final result depends on the choice of the initial point, which is easy to fall into the local optimum. In addition, it is relatively sensitive to noise and abnormal pointsSinaga and Yang (2020).

Algorithm 5 k-means Algorithm

```
Input:
    Training set: $T = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)\}$,
    Number of clusters: $K$.
Output:
    Cluster division: $C = C_1, C_2, ..., C_k$
    Steps:
Repeat the iteration as follows until the algorithm converges:
    Initialization stage: take $C_k = \phi, k = 1, 2, ..., K$
    Divide the stage: make $i = 1, 2, ..., m$
        The cluster class label for calculating $\boldsymbol{x_i}$ is as follows:
            $$\lambda_i = arg \min_k \parallel \boldsymbol{x_i} - \boldsymbol{\mu_k} \parallel_2, k \in \{1, 2, ..., K\} \tag{10}$$
        Then divide the sample $\boldsymbol{x_i}$ into the corresponding cluster:
            $$C_{\lambda_i} = C_{\lambda_i} \bigcup \{\boldsymbol{x_i}\} \tag{11}$$
    Recalculation stage: calculate $\boldsymbol{\mu_k}^*$
            $$\boldsymbol{\mu_k}^* = \frac{1}{|C_k|} \sum_{\boldsymbol{x_i} \in C_k} \boldsymbol{x_i} \tag{12}$$
    Termination condition:
        If there is $\boldsymbol{\mu_k}^* = \boldsymbol{\mu_k}$ for all $k \in 1, 2, ..., K$, the algorithm converges and the iteration is terminated.
        Otherwise, reassign $\boldsymbol{\mu_k} = \boldsymbol{\mu_k}^*$
```

3.5.2. Application of k-means in financial market

In the stock market, Shin et al. used fuzzy K-means cluster analysis to segment customers in two trading modes Shin and Sohn (2004). Based on the similarity measurement and item clustering, Zhu et al. proposed a financial risk indicator system based on the K-means clustering algorithm, performed indicator selection and data processing, and constructed a financial risk early warning model based on the K-means clustering algorithm to classify and optimize financial risks Zhu and Liu (2021). Brusco et al. proposed a variable selection heuristic method based on non-hierarchical K-means clustering analysis with adjusted Rand index, which is used to measure cluster recovery and applied to cluster analysis of real-world financial service data Brusco and Cradit (2001). Rokade et al. applied the K-means clustering algorithm, which uses unsupervised learning to cluster data sets, to analyze the stock market of listed companies in India, trying to isolate companies that exhibit abnormally high growth Rokade et al. (2016). Roshan et al. proposed a prediction model that combines K-means clustering, wavelet transform and support vector machines to predict financial markets to support decision-making in real-world transactions Roshan et al. (2016). Based on the sliding window truncating the stock price sequence into a chart, Yung et al. used the K-means algorithm to cluster the chart to form a chart form to construct a model for predicting stock trends Yung-Piao et al. (2014). Desokey et al. combined genetic algorithm and k-means to develop a clustering algorithm to try to enhance the previous intelligent framework that uses mixed social media and market data for stock forecast analysis Desokey et al. (2017). Based on the idea of clustering outlier detection, Zhu et al. proposed an improved K-MEANS clustering algorithm to detect outliers in financial time series Zhu and Che (2014). Based on the existing two-stage fusion model in the literature, Xu et al. proposed a new prediction model that uses the k-means clustering method to cluster commonly used technical indicators to predict the closing price of the stock market Xu et al. (2020). After investigating the quality of voluntary disclosure of environmental policies by financial intermediaries, and analyzing data from the Global Reporting Initiative, Alberici et al. identified four parts of financial intermediaries through k-means cluster analysis Alberici and Querci (2015). Li et al. analyzed the investment efficiency of smart investment based on K-means clustering analysis and data mining technology to conduct research on CMB Capricorn Intelligence Li et al. (2020). Took the above 17-year high-frequency data of the Stock Composite Index as an example. Huang et al. combined the multifractal algorithm with k-mean clustering to propose a new multifractal clustering model Huang and Tang (2021). Dai et al. used the traditional K-means algorithm to select the initial clustering center, and built a data mining system based on the improved algorithm to manage and supervise financial industry projects Dai (2021). Based on the information of news articles, Seong et al. proposed a method to reflect the heterogeneity and search for a group of homogeneous companies with high correlation, using K-means clustering and multi-core learning techniques to predict stock price trends Seong and Nam (2021).

3.6. Expectation-maximization algorithm

The EM algorithm is an iterative optimization strategy, because each iteration in its calculation method is divided into two steps, one of which is the desired step and the other is the maximum step. The EM algorithm is affected by the lack of thought. It is originally to solve the problem of parameter estimation in the case of missing data. The algorithm foundation and the effectiveness of convergence are explained in detail Dempster et al. (1977).

3.6.1. Introduction to Expectation-maximization algorithm

The EM algorithm has developed rapidly in recent decades. Its basic idea is relatively simple. First, estimate the value of the model parameters based on the observation data that has been given; then estimate the missing data based on the parameter values estimated in the previous step. According to the estimated missing data plus the previously observed data, re-estimate the parameter value, and then iterate repeatedly until it finally converges, and the iteration ends. With the development of theory, the EM algorithm has not only been used to deal with the problem of missing data, but with this kind of thinking, it can deal with a wider range of problems. Sometimes the missing data is not really missing, but a strategy adopted to simplify the problem. At this time, the EM algorithm is called a data addition technology, and the added data is usually called "potential data". Complex problems are introduced by introducing appropriate potential data can effectively solve our problems Abdalmageed et al. (2003). The EM algorithm is described as follows AlgorithmAlg$_e m$:

3.6.2. Application of Expectation-maximization algorithm in financial market

By using modified Zigzag technical indicators to segment data and discover topics, use expectation maximization to cluster, and use support vector machines to classify topics and predict accurate transaction parameters for identified, Ozorhan et al. proposed a method to predict the short-term trend of financial time series data in the foreign exchange market Özorhan et al. (2018). By using a random version of the expectation maximization algorithm to infer the parameters, Centanni et al. proposed a calculation simulation method applied to hedging strategies and derivative prices Centanni and Minozzo (2006). Chen et al. used the study of expectation maximization to detect suspicious transactions to prevent money laundering activities Chen et al. (2014b). Mazzarisi et al. proposed an expectation maximization algorithm for model estimation and latent variable inference, which can be used to predict the existence of future links and be applied to the e-MID inter-bank network Mazzarisi et al. (2020). Saidane et al. used dynamic switching in the hidden Markov model of mixed conditional heteroscedasticity factor analysis, and discussed the Viterbi algorithm in combination with the expectation maximization algorithm to iterative estimate the model parameters in the sense of maximum likelihood Saidane and Lavergne (2009). By constructing a joint maximum likelihood estimation of all model parameters through the expectation-maximization algorithm, Naf et al. proposed a mean-square-difference-mass-tail mixed distribution to model financial asset returns Nf et al. (2019). In response to financial disasters, Bernardi et al. developed a new type of flexible copula model, which relies on the expectation maximization algorithm and Markov switching generalized auto-regressive scores Dynamic changes Bernardi and Catania (2018). NG et al. proposed a stock analysis framework StockProF for quickly constructing stock investment portfolios, which used local anomaly factors and expectation maximization methods Ng and Khor (2016). Philip et al. proposed a clustering algorithm based on an improved hidden Markov model to predict the repayment behavior of customers of financial institutions that provide loans, which uses an expectation maximization algorithm to impose constraints Philip et al. (2018). Pei et al. proposed a 3D enhanced convolutional network based on the expectation maximization algorithm to extract time series information and solve the serious data imbalance problem Pei et al. (2020). Paolella et al. used a new two-stage expectation maximization algorithm to uniformly estimate all model parameters, which avoided most of the losses during the financial crisis and greatly improved the risk-adjusted return Paolella et al. (2019). Zoghi et al. used the expectation maximization algorithm to estimate the parameters of this distribution to optimize the use of value-at-risk and conditions as risk metrics Zoghi et al. (2021).

3.7. Ensemble learning

In data mining classification algorithms, the general goal is to learn a stable model that performs well in all aspects, but the actual situation is often not so ideal, and sometimes only multiple weak classification models with preferences can be obtained. Ensemble learning is to combine multiple weak classification models here in order to obtain a better and more comprehensive strong classification model. The underlying idea of ensemble learning is that even if a certain weak classifier gets a wrong prediction, other weak classifiers can also correct the error Webb and Zheng (2004); Dong et al. (2019).

3.7.1. Introduction to ensemble learning

Boosting is a commonly used method in the ensemble,, which is a machine learning algorithm that can be used to reduce the bias in supervised learning. The Boosting method learns a series of weak classifiers and combines them into a strong classifier. The representative of Boosting is the AdaBoost algorithm. The basic idea is to assign equal weight to each training example at the beginning of training, and then use the algorithm to train the training set for *t* rounds. After each training, the training examples that fail to train will be given a larger weight, that is, let the learning algorithm pays more attention to the wrong samples after each learning, to obtain multiple prediction functions Rudin et al. (2004). The AdaBoost algorithm is described as follows Algorithm $\mathrm{Alg}_e l$:

3.7.2. Application of ensemble learning in financial market

Combining swarm intelligence and technical analysis, Wang et al. proposed a machine learning technology that combines deep learning algorithms and ensemble learning methods for financial market prediction Wang et al. (2018). Sun et al. proposed a hybrid ensemble learning method combining the AdaBoost algorithm and long-term short-term memory network to predict financial time series Sun et al. (2018b). Ferreira et al. considered five single classifiers, three ensemble classifiers using decision trees as weak classifiers, and four ensemble classifiers that combined other eight classifiers to predict the future trend of financial asset prices Ferreira et al. (2020). Sreedharan et al. introduced the use of ensemble techniques (including majority voting (MV), random forest, and AdaBoost ensemble) to build more accurate predictive models to predict financial distress by combining the outputs of various classifiers Sreedharan et al. (2020). Combining the decision tree stump with the Adaboost algorithm, Zhang et al. proposed the DS_Adaboost algorithm to provide early warning of financial problems Zhang and Jiang (2017). Liu et al. proposed a model that combines the advantages of ensemble learning and deep learning to predict changes in stock prices Liu et al. (2019). Gonzalez et al. proposed the design of an integrated system based on genetic algorithms to predict the weekly price trend of the Sao Paulo Stock Exchange index Gonzalez et al. (2015). Ekinci et al. used attribute-based integrated learning methods and two instance-based integrated learning methods to improve the prediction accuracy of traditional machine learning models for bank failure prediction Ekinci and Erdal (2016). Li et al. proposed a SVM ensemble based on Choquet points for financial distress prediction, in which a new training set was generated using the Bagging algorithm Li et al. (2015). Julia et al. proposed and evaluated some algorithmic algotrading strategies based on artificial neural network integration to support the decision-making of stock market investors Julia et al. (2018). In order to understand which models should be combined and how to effectively deal with features related to different aspects in different models, Cagliero et al. studied the use of ensemble methods to combine faceted classification models to support stock trading Cagliero et al. (2020).

Carta et al. proposed an ensemble learning model to minimize these problems without using annotations to learn reinforcement learning strategies, but to learn how to maximize the reward function in the training phase Carta et al. (2021). Zhu, etc. applied bagging, boosting and random subspace three integrated machine learning methods, and RS-boosting and multiboosting two ensemble machine learning methods integrated machine learning methods to predict the credit risk of small and medium-sized enterprises in supply chain finance Zhu et al. (2016). Liang et al. introduced a classifier ensemble method that aims to reduce the cost of misclassification. It uses a consensus voting method to combine the outputs generated by multiple classifiers to find the final prediction results to study financial distress Liang et al. (2017). Hsu et al. proposed a hybrid ensemble learning prediction mechanism, which comes from the extreme learning machine algorithm of different ensemble strategies such as data diversity, parameter diversity, nuclear diversity and preprocessing diversity, to evaluate the company's governance status Hsu and Lin (2014). Kilimci et al. used a variety of learning algorithm integration to improve the performance of the classification system, and proposed to estimate the direction of the Bist100 index through financial sentiment analysis Kilimci (2019). Based on the development and adoption of a new method of re-sampling financial sequences, Borges et al. proposed an ensemble learning system based on machine learning, aiming to create an investment strategy that can be traded in the cryptocurrency trading market Borges and Neves (2020). Based on neural network regression integration, support vector regression integration, enhanced regression tree, and random forest regression, Weng et al. developed an AI financial expert system to predict short-term stock prices Weng et al. (2018). Aljawazneh et al. compared the performance of three deep learning methods (long and short-term memory, deep belief network and 6-layer multi-layer perceptron model) with three bagging ensemble classifiers (random forest, support vector machine and K-Nearest Neighbor) and two boosting integrated classifiers (Adaptive Boosting and Extreme Gradient Boosting) for the prediction of the company's financial failure Aljawazneh et al. (2021). Tsai et al. developed a new hybrid financial distress model based on the combination of clustering technology and classifier integration by using two clustering technologies of self-organizing mapping and K-means, as well as three classification technologies of logistic regression, multilayer perceptron neural network and decision tree Tsai (2014). Based on the deep learning classifier, He et al. proposed a new hybrid ensemble model to improve the performance of predicting default in emerging financial markets He and Fan (2021).

4. Discussion

In the past few decades, people have become more and more interested in using data mining algorithms to analyze and predict multiple financial markets, especially in some areas of financial market segmentation, such as stock prices, exchange rates, financial distress, commodity price forecasts, etc. Although data mining technology has been widely used in financial data analysis, due to the high complexity, rapid growth and large amount of financial market data, effective analysis of financial data still poses greater challenges. In this section, we have formulated some problems that have not been well solved by current work and need to be further studied in the future.

Although there is a large amount of literature on time series analysis, most existing methods do not consider that financial time series are a special data stream. As a special case of data flow, financial time series often have conceptual drifts. Concept drift is a very big challenge to current commonly used data mining models, and it is also one of the future research directions.

There are many factors that affect the financial market. Any political, real economy, medical emergencies, etc. will have a huge impact on the financial market. Even some small events that people rarely pay attention to may have an impact on financial markets. Current data mining algorithms mainly focus on a single data source. How to integrate multiple data sources to conduct data mining and analysis from different angles is one of the challenges and research directions in the future.

5. Conclusions

This work reviews data mining algorithms used to solve financial market problems. The survey focuses on the application of data mining algorithms in the financial market. Therefore, in this review, different data mining algorithms are classified, and more than 100 documents dealing with financial data preprocessing, classification and clustering, future trend prediction and financial information mining are investigated and discussed respectively.

One of the important contributions of this work is to discuss the basic concepts of all reviewed data mining algorithms. The basic principles of these data mining algorithms are elaborated and discussed, and the basic main steps of the algorithms are described. This method can be used to guide students, investors and practitioners to construct intelligent financial forecasting methods based on data mining algorithms. Another important contribution of this work is to review the latest literature on the subject according to the classification of data mining algorithms. Although other comments have been published recently, they differ from this work in terms of scope or the main research investigated. In addition, we also discuss the main challenges and future directions in this research field.

It is worth mentioning that this review has some limitations. It may be missing related articles that are not indexed in the selected database. In addition, some data mining methods are not included in this work, especially deep learning, which is very popular recently. One of our work in the future is to summarize the application of deep learning in the financial market.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (grant number 62172158) and the Project of Hunan Institute of Technology (grant number HQ20004).

Conflict of interest

The authors declare no conflict of interest.

References

Abdalmageed W, Elosery A, Smith CE (2003) Non-parametric expectation maximization: a learning automata approach. In *IEEE International Conference on Systems*, 2003.

Agrawal L, Adane D (2021) Improved decision tree model for prediction in equity market using heterogeneous data. *IETE J Res*, 1–10.

Ahn JJ, Oh KJ, Kim TY, et al. (2011) Usefulness of support vector machine to develop an early warning system for financial crisis. *Expert Syst Appl* 38: 2966–2973.

Alberici A, Querci F (2015) The quality of disclosures on environmental policy: The profile of financial intermediaries. *Corp Soc Resp Env Ma* 23: 283–296.

Aljawazneh H, Mora AM, Garcia-Sanchez P, et al. (2021) Comparing the performance of deep learning methods to predict companies' financial failure. *IEEE Access* 9: 97010–97038.

Atsalakis GS, Valavanis KP (2009) Surveying stock market forecasting techniques – part II: Soft computing methods. *Expert Syst Appl* 36: 5932–5941.

Javed Awan M, Mohd Rahim MS, Nobanee H, et al. (2021) Social media and stock market prediction: A big data approach. *Comput Mater Con* 67: 2569–2583.

Barboza F, Kimura H, Altman E (2017) Machine learning models and bankruptcy prediction. *Expert Syst Appl* 83: 405–417.

Bernardi M, Catania L (2018) Switching generalized autoregressive score copula models with application to systemic risk. *J Appl Econometrics* 34: 43–65.

Bielza C, Larranaga P (2014) Discrete bayesian network classifiers. *ACM Comput Surv* 47: 1–43.

Bishop CM (2006) *Pattern Recognition and Machine Learning*. Springer New York, 2006.

Borges TA, Neves RF (2020) Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods. *Appl Soft Comput* 90: 106187.

Braun B (2018) Central banking and the infrastructural power of finance: the case of ECB support for repo and securitization markets. *Socio-Econ Rev* 18: 395–418.

Brusco MJ, Cradit JD (2001) A variable-selection heuristic for k-means clustering. *Psychometrika* 66: 249–270.

Burges CJ (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 2: 121–167.

Bustos O, Pomares-Quimbaya A (2020) Stock market movement forecast: A systematic review. *Expert Syst Appl* 156: 113464.

Cagliero L, Garza P, Attanasio G, et al. (2020) Training ensembles of faceted classification models for quantitative stock trading. *Computing* 102: 1213–1225.

Cao LJ, Tay FEH (2003) Support vector machine with adaptive parameters in financial time series forecasting. *IEEE T Neural Networ* 14: 1506–1518.

Carpinteiro OA, Leite JP, Pinheiro CA, et al. (2011) Forecasting models for prediction in time series. *Artif Intell Rev* 38: 163–171.

Carta S, Ferreira A, Podda AS, et al. Multi-DQN: An ensemble of deep q-learning agents for stock market forecasting. *Expert Syst Appl* 164: 113820.

Cavalcante RC, Brasileiro RC, Souza VL, et al. Computational intelligence and financial markets: A survey and future directions. *Expert Syst Appl* 55: 194–211.

Celebi ME, Kingravi HA, Vela PA (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst Appl* 40: 200–210.

Centanni S, Minozzo M (2006) Estimation and filtering by reversible jump MCMC for a doubly stochastic poisson model for ultra-high-frequency financial data. *Stat Model* 6: 97–118.

Chen AS, Leung MT, Pan S (2019) Financial hedging in energy market by cross-learning machines. *Neural Comput Appl* 32: 10321–10335.

Chen HL, Liu DY, Yang B, et al. (2011) An adaptive fuzzy k-nearest neighbor method based on parallel particle swarm optimization for bankruptcy prediction. In *Adv Knowl Discovery Data Min*, 249–264. Springer Berlin Heidelberg, 2011.

Chen MY (2011) Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Syst Appl* 38: 11261–11272.

Chen S (2019) An effective going concern prediction model for the sustainability of enterprises and capital market development. *Appl Econ* 51: 3376–3388.

Jin C, De-Lin L, Fen-Xiang M (2014) An improved ID3 decision tree algorithm. *Adv Mater Res* 962-965: 2842–2847.

Chen Y, Hao Y (2017) A feature weighted support vector machine and k-nearest neighbor algorithm for stock market indices prediction. *Expert Syst Appl* 80: 340–355.

Chen Z, Nazir A, Teoh EN, et al. Exploration of the effectiveness of expectation maximization algorithm for suspicious transaction detection in anti-money laundering. In *2014 IEEE Conference on Open Systems (ICOS)*. IEEE.

Cheng SH (2014) Predicting stock returns by decision tree combining neural network. *Lect Notes Artif Int* 8398: 352–360.

Cheng CH, Chan CP, Sheu YJ (2019) A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Eng Appl Artif Intel* 81: 283–299.

Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20: 273–297.

Dai W (2021) Development and supervision of robo-advisors under digital financial inclusion in complex systems. *Complexity* 2021: 1–12.

Daugaard D Emerging new themes in environmental, social and governance investing: a systematic literature review. *Account Financ* 60: 1501–1530.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via theEMAlgorithm. *J Royal Stat Soc* 39: 1–22.

Deng S, Wang C, Wang M, et al. (2019) A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of china stock market. *Appl Soft Comput* 83: 105652.

Desokey EN, Badr A, Hegazy AF Enhancing stock prediction clustering using k-means with genetic algorithm. In *2017 13th International Computer Engineering Conference (ICENCO)*. IEEE.

Dong X, Yu Z, Cao W, et al. (2019) A survey on ensemble learning. *Front Comput Sci* 14: 241–258.

Ekinci A, Erdal HI (2016) Forecasting bank failure: Base learners, ensembles and hybrid ensembles. *Comput Econ* 49: 677–686.

Farid S, Tashfeen R, Mohsan T, et al. (2020) Forecasting stock prices using a data mining method: Evidence from emerging market. *Int J Financ Econ*.

Ferreira FGDC, Gandomi AH, Cardoso RTN (2020) Financial time-series analysis of brazilian stock market using machine learning. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE.

Ferreira LEB, Barddal JP, Gomes HM, et al. (2017) Improving credit risk prediction in online peer-to-peer (p2p) lending using imbalanced learning techniques. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Fields D Constructing a new asset class: Property-led financial accumulation after the crisis. *Econ Geogr* 94: 118–140.

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. *Mach Learn* 29: 131–163.

Gamage P (2016) New development: Leveraging 'big data' analytics in the public sector. *Public Money Manage* 36: 385–390.

García S, Fernández A, Herrera F (2009) Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. *Appl Soft Comput* 9: 1304–1314.

Garcia-Almanza AL, Tsang EP (2006) The repository method for chance discovery in financial forecasting, In *International Conference on Knowledge-based Intelligent Information and Engineering Systems*.

Gonzalez RT, Padilha CA, Barone DAC (2015) Ensemble system based on genetic algorithm for stock market forecasting. In *2015 IEEE Congress on Evolutionary Computation (CEC)*. IEEE.

Gou J, Ma H, Ou W, et al. (2019) A generalized mean distance-based k-nearest neighbor classifier. *Expert Syst Appl* 115: 356–372.

Goyal K, Kumar S (2020) Financial literacy: A systematic review and bibliometric analysis. *Int J Consum Stud* 45: 80–105.

Guo S, He H, Huang X (2019) A multi-stage self-adaptive classifier ensemble model with application in credit scoring. *IEEE Access* 7: 78549–78559.

Han J, Pei J, Kamber M (2000) *Data Mining: Concepts and Techniques*.

Han J, Cheng H, Xin D, et al. (2007) Frequent pattern mining: current status and future directions. *Data Min Knowl Discovery* 15: 55–86.

He H, Fan Y (2021) A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. *Expert Syst Appl* 176: 114899.

He S, Zheng J, Lin J, et al. (2020) Classification-based fraud detection for payment marketing and promotion. *Comput Syst Sci Eng* 35: 141–149.

Howe D, Costanzo M, Fey P, et al. (2008) The future of biocuration. *Nature* 455: 47–50.

Hssina B, Merbouha A, Ezzikouri H, et al. (2014) A comparative study of decision tree ID3 and c4.5. *Int J Adv Comput Sci Appl* 4.

Hsu YS, Lin SJ (2014) An emerging hybrid mechanism for information disclosure forecasting. *Int J Mach Learn Cybern* 7: 943–952.

Huang C, Gao F, Jiang H (2014) Combination of biorthogonal wavelet hybrid kernel OCSVM with feature weighted approach based on EVA and GRA in financial distress prediction. *Math Probl Eng* 2014: 1–12.

Huang Q, Wang T, Tao D, et al. (2015) Biclustering learning of trading rules. *IEEE T Cybern* 45: 2287–2298.

Huang X, Tang H (2021) Measuring multi-volatility states of financial markets based on multifractal clustering model. *J Forecast*.

Iqbal R, Doctor F, More B, et al. (2020) Big data analytics: Computational intelligence techniques and application areas. *Technol Forecast Soc* 153: 119253.

Jagadish HV, Gehrke J, Labrinidis A, et al. (2014) Big data and its technical challenges. *Commun ACM* 57: 86–94.

Rutkowski L, Jaworski M, Pietruczuk L, et al. (2014) The cart decision tree for mining data streams. *Infor Sci*.

Julia D, Pereira A, Silva RE (2018) Designing financial strategies based on artificial neural networks ensembles for stock markets. 1–8.

Kanhere P, Khanuja HK (2015) A methodology for outlier detection in audit logs for financial transactions. In *2015 International Conference on Computing Communication Control and Automation*. IEEE.

Kercheval AN, Zhang Y (2015) Modelling high-frequency limit order book dynamics with support vector machines. *Quant Financ* 15: 1315–1329.

Kewat P, Sharma R, Singh U, et al. (2017) Support vector machines through financial time series forecasting. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. IEEE.

Kilimci ZH (2019) Borsa tahmini için derin topluluk modellleri (DTM) ile finansal duygu analizi. *Gazi niversitesi Mhendislik-Mimarlık Fakltesi Dergisi*.

Kim SY, Upneja A (2021) Majority voting ensemble with a decision trees for business failure prediction during economic downturns. *J Innovation Knowl* 6: 112–123.

Kim YJ, Baik B, Cho S (2016) Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Syst Appl* 62: 32–43.

Kirkos E, Spathis C, Manolopoulos Y (2007) Data mining techniques for the detection of fraudulent financial statements. *Expert Syst Appl* 32: 995–1003.

Kotsiantis SB (2011) Decision trees: a recent overview. *Artif Intell Rev* 39: 261–283.

Kum HC, Ahalt S, Carsey TM (2011) Dealing with data: Governments records. *Science* 332: 1263–1263.

Kumar DA, Murugan S (2013) Performance analysis of indian stock market index using neural network time series model. In *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*. IEEE.

Lee I (2017) Big data: Dimensions, evolution, impacts, and challenges. *Bus Horizons* 60: 293–303.

Lee TK, Cho JH, Kwon DS, et al. (2019) Global stock market investment strategies based on financial network indicators using machine learning techniques. *Expert Syst Appl* 117: 228–242.

Li H, Sun J, Sun BL (2009) Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. *Expert Syst Appl* 36: 643–659.

Li L, Wang J, Li X (2020) Efficiency analysis of machine learning intelligent investment based on k-means algorithm. *IEEE Access* 8: 147463–147470.

Li ST, Ho HF (2009) Predicting financial activity with evolutionary fuzzy case-based reasoning. *Expert Syst Appl* 36: 411–422.

Li T, Li J, Liu Z, et al. (2018) Differentially private naive bayes learning over multiple data sources. *Inf Sci* 444: 89–104.

Li X, Wang F, Chen X (2015) Support vector machine ensemble based on choquet integral for financial distress prediction. *Int J Pattern Recognit Artif Intell* 29: 1550016.

Liang D, Tsai CF, Dai AJ, et al. (2017) A novel classifier ensemble approach for financial distress prediction. *Knowl Inf Syst* 54: 437–462.

Liao SH, Chu PH, Hsiao PY (2012) Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Syst Appl* 39: 11303–11311.

Lin A, Shang P, Feng G, et al. (2012) APPLICATION OF EMPIRICAL MODE DECOMPOSITION COMBINED WITH k-NEAREST NEIGHBORS APPROACH IN FINANCIAL TIME SERIES FORECASTING. *Fluct Noise Lett* 11: 1250018.

Lin CS, Chiu SH, Lin TY (2012) Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Econ Model* 29: 2583–2590.

Lin G, Lin A, Cao J (2021) Multidimensional KNN algorithm based on EEMD and complexity measures in financial time series forecasting. *Expert Syst Appl* 168: 114443.

Liu J, Lin CMM, Chao F (2019) Gradient boost with convolution neural network for stock forecast. In *Adv Intell Syst Comput*, 155–165.

Liu M, Luo K, Zhang J, et al. (2021) A stock selection algorithm hybridizing grey wolf optimizer and support vector regression. *Expert Syst Appl* 179: 115078.

Liu W, Zhao J, Wang D (2021) Data mining for energy systems: Review and prospect. *WIREs Data Min Knowl Discovery* 11.

Jan CL (2018) An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from taiwan. *Sustainability* 10: 513.

Loukeris N, Eleftheriadis I, Livanis E (2013) A novel approach on hybrid support vector machines into optimal portfolio selection. In *IEEE Int Symposium Signal Proc Inf TechnoL*. IEEE.

Luintel KB, Khan M, Leon-Gonzalez R, et al. (2016) Financial development, structure and growth: New data, method and results. *J Int Financ Mark Inst Money* 43: 95–112.

Luo B, Lin Z (2011) A decision tree model for herd behavior and empirical evidence from the online p2p lending market. *Inf Syst e-Bus Manage* 11: 141–160.

Ma Y, Xu B, Xu X (2017) Real estate confidence index based on real estate news. *Emerg Mark Financ Tr* 54: 747–760.

Malliaris AG, Malliaris M (2015) What drives gold returns? a decision tree analysis. *Financ Res Lett* 13: 45–53.

Mazzarisi P, Barucca P, Lillo F, et al. (2020) A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market. *Eur J Oper Res* 281: 50–65.

Mir-Juli M, Fiol-Roig G, Isern-Dey AP (2010) Decision trees in stock market analysis: Construction and validation. In *Trends Applied Intelligent Systems-international Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, 2010.

Muja M, Lowe DG (2014) Scalable nearest neighbor algorithms for high dimensional data. *IEEE T Pattern Anal* 36: 2227–2240.

Naranjo R, Santos M (2019) A fuzzy decision system for money investment in stock markets based on fuzzy candlesticks pattern recognition. *Expert Syst Appl* 133: 34–48.

Nardo M, Petracco-Giudici M, Naltsidis, M (2015) WALKING DOWN WALL STREET WITH a TABLET: A SURVEY OF STOCK MARKET PREDICTIONS USING THE WEB. *J Econ Surv* 30: 356–369.

Al Nasseri A, Tucker A, de Cesare S (2015) Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Syst Appl* 42: 9192–9210.

Nassirtoussi AK, Aghabozorgi S, Wah TY, et al. (2014) Text mining for market prediction: A systematic review. *Expert Syst Appl* 41: 7653–7670.

Nf J, Paolella MS, Polak P (2019) Heterogeneous tail generalized COMFORT modeling via cholesky decomposition. *J Multivariate Anal* 172: 84–106.

Ng A, Jordan M (2002) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002. URL `https://proceedings.neurips.cc/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf`.

Ng KH, Khor KC (2016) StockProF: a stock profiling framework using data mining approaches. *Inf Syst e-Bus Manage* 15: 139–158.

Nie CX (2020) A network-based method for detecting critical events of correlation dynamics in financial markets. *EPL (Europhys Lett)* 131: 50001.

Ohana JJ, Ohana S, Benhamou E, et al. (2021) Explainable AI (XAI) models applied to the multi-agent environment of financial markets. In *Explainable and Transparent AI and Multi-Agent Systems*, pages 189–207. Springer International Publishing, 2021.

Olson DL (2006) Data mining in business services. *Serv Bus* 1: 181–193.

Oussous A, Benjelloun FZ, Lahcen AA, et al. (2018) Big data technologies: A survey. *J King Saud University - Comput Inf Sci* 30: 431–448.

Pan I, Bester D (2018) Fuzzy bayesian learning. *IEEE T Fuzzy Syst* 26: 1719–1731.

Paolella MS, Polak P, Walker PS (2019) Regime switching dynamic correlations for asymmetric and fat-tailed conditional returns. *J Econometrics* 213: 493–515.

Patrizio A (2018) Idc: Expect 175 zettabytes of data worldwide by 2025. https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html.

Pei S, Shen T, Wang X, et al. (2020) 3dacn: 3d augmented convolutional network for time series data. *Inf Sci* 513: 17–29.

Peng Y, Wang G, Kou G, et al. (2011) An empirical study of classification algorithm evaluation for financial risk prediction. *Appl Soft Comput* 11: 2906–2915.

Philip DJ, Sudarsanam N, Ravindran B (2018) Improved insights on financial health through partially constrained hidden markov model clustering on loan repayment data. *ACM SIGMIS Database DATABASE Adv Inf Syst* 49: 98–113.

Provost F, Fawcett T (2013) Data science and its relationship to big data and data-driven decision making. *Big Data* 1: 51–59.

Qian B, Rasheed K (2006) Stock market prediction with multiple classifiers. *Appl Intell* 26: 25–33.

Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1: 81–106.

Raudys Š (2000) How good are support vector machines? *Neural Networks* 13: 17–19.

Rokade A, Malhotra A, Wanchoo A (2016) Enhancing portfolio returns by identifying high growth companies in indian stock market using artificial intelligence. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE.

Rosati R, Romeo L, Goday CA (2020) Machine learning in capital markets: Decision support system for outcome analysis. *IEEE Access* 8: 109080–109091.

Roshan WDS, Gopura RARC, Jayasekara AGB, et al. (2016) Financial market forecasting by integrating wavelet transform and k-means clustering with support vector machine. In *International Conference on Artificial Life and Robotics*, 2016.

Roychowdhury S, Shroff N, Verdi RS (2019) The effects of financial reporting and disclosure on corporate investment: A review. *J Account Econ* 68: 101246.

Rudin C, Daubechies I, Schapire RE, et al. (2004) The dynamics of adaboost: Cyclic behavior and convergence of margins. *J Mach Learn Res* 5: 1557–1595.

Ryans JP (2020) Textual classification of SEC comment letters. *Rev Account Stud* 26: 37–80.

Saidane M, Lavergne C (2009) Optimal prediction with conditionally heteroskedastic factor analysed hidden markov models. *Comput Econ* 34: 323–364.

Salzberg SL (1994) C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Mach Learn* 16: 235–240.

Samworth RJ (2012) Optimal weighted nearest neighbour classifiers. *Annal Stat* 40.

Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news. *ACM T Inf Syst* 27: 1–19.

Seong N, Nam K (2021) Predicting stock movements based on financial news with segmentation. *Expert Syst Appl* 164: 113988.

Shamim S, Zeng J, Shariq SM, et al. (2019) Role of big data management in enhancing big data decision-making capability and quality among chinese firms: A dynamic capabilities view. *Inform Manage* 56: 103135.

Shin HW, Sohn SY (2004) Segmentation of stock trading customers according to potential value. *Expert Syst Appl* 27: 27–33.

Si YW, Yin J (2013) OBST-based segmentation approach to financial time series. *Eng Appl Artif Intel* 26: 2581–2596.

Sinaga KP, Yang MS (2020) Unsupervised k-means clustering algorithm. *IEEE Access* 8: 80716–80727.

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14: 199–222.

Soni S (2011) Applications of anns in stock market prediction: A survey. *Int J Comput Sci Eng Technol* 2: 71–83.

Sreedharan M, Khedr AM, El Bannany M (2020) A comparative analysis of machine learning classifiers and ensemble techniques in financial distress prediction. In *2020 17th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 653–657.

Sun H, Rong W, Zhang J, et al. (2017) Stacked denoising autoencoder based stock market trend prediction via k-nearest neighbour data selection. In *International Conference on Neural Information Processing*. Springer, 882–892.

Sun J, Lang J, Fujita H, et al. (2018a) Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Inf Sci* 425: 76–91.

Sun J, Li H, Fujita H, et al. (2020) Class-imbalanced dynamic financial distress prediction based on adaboost-SVM ensemble combined with SMOTE and time weighting. *Inform Fusion* 54: 128–144.

Sun SL, Wei YJ, Wang SY (2018b) AdaBoost-LSTM ensemble learning for financial time series forecasting. In *International Conference on Computational Science*. Springer, 590–597.

Talebi H, Hoang W, Gavrilova ML (2014) Multi-scale foreign exchange rates ensemble for classification of trends in forex market. *Proc Comput Sci* 29: 2065–2075.

Tang L, Pan PH, Yao YY (2018a) EPAK: A computational intelligence model for 2-level prediction of stock indices. *Int J Comput Commun* 13: 268–279.

Tang XB, Liu GC, Yang J, et al. (2018b) Knowledge-based financial statement fraud detection system: based on an ontology and a decision tree. *Knowl Organ* 45: 205–219.

Tsai CF (2014) Combining cluster analysis with classifier ensembles to predict financial distress. *Inform Fusion* 16: 46–58.

Tsai CF, Chiou YJ (2009) Earnings management prediction: A pilot study of combining neural networks and decision trees. *Expert Syst Appl* 36: 7183–7191.

Vaghela VB, Vandra KH, Modi NK (2014) Mr-mnbc: Maxrel based feature selection for the multi-relational nave bayesian classifier. In *Nirma University International Conference on Engineering*, 1–9.

Wang B, Huang H, Wang X (2011a) A support vector machine based MSM model for financial short-term volatility forecasting. *Neural Comput Appl* 22: 21–28.

Wang JZ, Wang JJ, Zhang ZG, et al. (2011b) Forecasting stock indices with back propagation neural network. *Expert Syst Appl* 38: 14346–14355.

Wang L, Zhu J (2008) Financial market forecasting using a two-step kernel learning method for the support vector regression. *Ann Oper Res* 174: 103–120.

Wang Q, Xu W, Zheng H (2018) Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing* 299: 51–61.

Webb GI, Zheng Z (2004) Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. *IEEE T Knowl Data En* 16: 980–991.

Weng B, Lu L, Wang X, et al. (2018) Predicting short-term stock prices using ensemble methods and online data sources. *Expert Syst Appl* 112: 258–273.

Wu XD, Kumar V, Quinlan JR, et al. (2007) Top 10 algorithms in data mining. *Knowl Inf Syst* 14: 1–37.

Xing FZ, Cambria E, Welsch RE (2017) Natural language based financial forecasting: a survey. *Artif Intell Rev* 50: 49–73.

Xu Y, Yang C, Peng S, et al. (2020) A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning. *Appl Intell* 50: 3852–3867.

Yan L, Bai B (2016) Correlated industries mining for chinese financial news based on LDA trained with research reports. In *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 131–135.

Yang R, Yu L, Zhao Y, et al. (2020) Big data analytics for financial market volatility forecast based on support vector machine. *Int J Inf Manag* 50: 452–462.

Yeo B, Grant D (2018) Predicting service industry performance using decision tree analysis. *Int J Inf Manag* 38: 288–300.

Yoo PD, Kim MH, Jan T (2005) Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC06)*. IEEE. 2: 835–841.

Zhang Y, Yu G, Jin ZQ (2013) Violations detection of listed companies based on decision tree and k-nearest neighbor. In *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings*, 1671–1676.

Wu KP, Wu YP, Lee HM (2014) Stock trend prediction by using k-means and aprioriall algorithm for sequential chart pattern mining. *J Inf Sci Eng* 30: 653–667.

Zemke S (1999) Nonlinear index prediction. *Physica A* 269: 177–183.

Chenggang Zhang and Jingqing Jiang. A financial early warning algorithm based on ensemble learning. In *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*. IEEE, sep 2017. doi: 10.1109/ciapp.2017.8167192.

Zhang H, Li SF (2010) Forecasting volatility in financial markets. *Key Eng Mater* 439: 679–682.

Zhang JL, Härdle WK (2010) The bayesian additive classification tree applied to credit risk modelling. *Comput Stat Data An* 54: 1197–1205.

Zhang N, Lin A, Shang P (2017) Multidimensionalk-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A* 477: 161–173.

Zhao QJ, SunQ, Che WG (2014) The application of bayesian discrimination in the analysis on media sector stock. *Applied Mechanics and Materials* 488: 1310–1313.

Zhao Y (2021) Sports enterprise marketing and financial risk management based on decision tree and data mining. *J Healthc Eng* 2021: 1–8.

Guo ZQ, Wang HQ, Liu Q (2012) Financial time series forecasting using LPP and SVM optimized by PSO. *Soft Comput* 17: 805–818.

Zhu X, Che WG (2014) Research of outliers in time series of stock prices based on improved k-means clustering algorithm. *Wit Trans Inf Commun Technol* 46: 633–641.

Zhu Y, Xie C, Wang GJ, et al. (2016) Comparison of individual, ensemble and integrated ensemble machine learning methods to predict china's SME credit risk in supply chain finance. *Neural Comput Appl* 28: 41–50.

Zhu Z, Liu N (2021) Early warning of financial risk based on k-means clustering algorithm. *Complexity* 2021: 1–12.

Zhuang Y, Xu Z, Tang Y (2015) A credit scoring model based on bayesian network and mutual information. In *2015 12th Web Information System and Application Conference (WISA).*

Mirsadeghpour Zoghi SM, Saneie M, Tohidi G, et al. (2021) The effect of underlying distribution of asset returns on efficiency in dea models. *Journal of Intelligent and Fuzzy Systems* 40: 10273–10283.

Özorhan MO, Toroslu İH, Şehitoğlu OT (2018) Short-term trend prediction in financial time series data. *Knowl Inf Syst* 61: 397–429.