10.70711/aitr.v2i5.5285

# Principles and Applications of Common Artificial Intelligence Algorithms in the Financial Field

**Xiaoliang Ji**

**ENAE Business School, Murcia (Spain) 30100**

***Abstract:*** This article summarizes common machine learning algorithms in the financial field, such as KNN, Kmeans, SVM, regression algorithms, etc., and explains their principles. It also elaborates on their common application scenarios in the financial field and advanced usage techniques such as hyperparameter optimization, Stacking, Boosting, etc.

***Keywords:*** Finance, Machine learning; Algorithms; KNN; Kmeans; SVM; Regression algorithms; Hyper parameter optimization Stacking; Boosting

## Introduction

There are many algorithms involved in artificial intelligence. For research purposes, this article summarizes common computer statistical learning methods (machine learning algorithms) in the financial field and explains their principles. Machine learning is the process of enabling computer systems to automatically analyze, learn, predict, and make decisions from data through algorithms. In machine learning, algorithms continuously learn from input data and generate models to interpret and predict the data. During this process, algorithms can be adjusted and optimized based on learning and reasoning from data to improve prediction accuracy and generalization ability.

## 1. Classification of Machine Learning Algorithms

Machine learning algorithms can be classified based on different learning methods and application scenarios. Common classification methods include supervised learning and unsupervised learning.

Supervised learning refers to machine learning conducted with labeled data. These labeled data include input and output variables used for training and optimizing the model.

Unsupervised learning refers to machine learning conducted without labeled data. Unsupervised learning algorithms discover structures

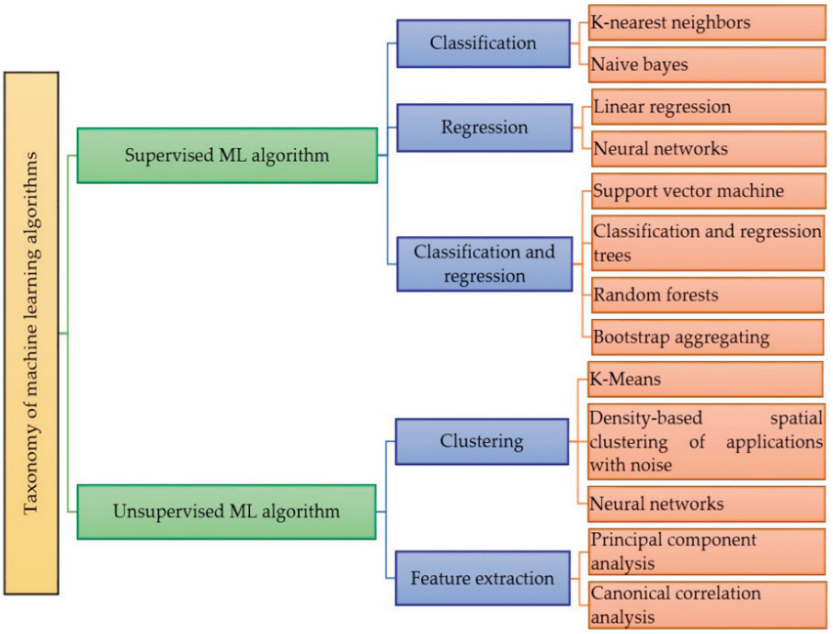


**Figure 1. the classic algorithms in supervised and unsupervised learning**

and patterns in input data through clustering, dimensionality reduction, and correlation analysis.

Reinforcement learning refers to machine learning through interaction with the environment. Reinforcement learning algorithms continuously improve models through the interaction between agents and the environment, and are not separately classified here.

The following figure summarizes the classic algorithms in supervised and unsupervised learning:

Various algorithms are widely used in various financial fields, including data mining, natural language processing, problem diagnosis, recommendation systems, etc. For example, data mining techniques can be used to process and analyze financial big data to discover hidden patterns and correlations in the data; In the field of natural language processing, tasks such as text classification, sentiment analysis, machine translation, and speech recognition can be achieved, replacing manual labor or providing personalized services for financial customers; In the field of problem diagnosis, machine learning algorithms can assist analysts or clients in making financial and business decisions; Various recommendation systems can provide financial product recommendations or classify and evaluate customers' credit ratings.

## 2. Analysis and Practice of Classic Algorithms

This article will provide a detailed introduction to eight commonly used machine learning algorithms: K-Nearest Neighbor (KNN) algorithm, K-means algorithm, decision tree algorithm, Naive Bayes algorithm, Support Vector Machine (SVM), regression algorithm, neural network algorithm, ensemble learning algorithm, and their principles and typical applications.

### 2.1 K-Nearest Neighbor Algorithm (KNN)

KNN is the most basic classification and regression algorithm. It classifies the input samples based on the categories or values of the k nearest neighbors of the training samples. In the financial field, KNN algorithm is often used to solve problems such as data classification, text classification, image classification, and preliminary stock trend analysis. The distance in KNN algorithm is often calculated by Euclidean distance or cosine of angle, but in big data scenarios, simplified operations such as Manhattan distance can be used to reduce computational complexity.

### 2.2 Kmeans algorithm

Kmeans is a clustering algorithm. The so-called clustering refers to the classification of a set of physical or abstract objects into multiple similar objects.

Data clustering algorithms can be divided into structural clustering and decentralized clustering. Structural clustering uses previously used classifiers for classification, while decentralized clustering determines all classifications at once. Structural clustering algorithms can perform calculations from top to bottom or from bottom to top. The bottom-up algorithm starts clustering from each object as a separate category and continuously integrates similar objects. The top-down algorithm classifies all objects as a whole and gradually breaks them down. In addition, there are distributed clustering algorithms that determine the categories to be generated at once. Distributed clustering algorithm is also a bottom-up algorithm.

Kmeans algorithm is the most basic decentralized clustering algorithm, and its steps are:

① Randomly select the number of clusters k;

② Generate k clusters arbitrarily, then determine the cluster centers, or directly generate k centers;

③ Determine the clustering center point for each point;

④ Calculate its new cluster center again;

⑤ Repeat the above steps until the convergence requirements are met (usually referring to the determined center point no longer changing).

### 2.3 Decision Tree Algorithm

The decision tree algorithm classifies or regresses input samples by constructing a tree like structure. The decision tree algorithm is intuitive and easy to understand, but it is susceptible to the effects of noisy data and overfitting. In the financial field, decision tree algorithms are often used to solve various classification and regression problems, such as credit card fraud detection, market forecasting, customer credit evaluation, expert systems, and various recommendation systems.

### 2.4 Naive Bayes Algorithm

Naive Bayes algorithm is a classification algorithm based on Bayes' theorem. It calculates the probability that the input sample belongs to a certain category and makes classification predictions based on this probability. Naive Bayes algorithm has the advantages of simplicity, ease of implementation, and high efficiency, especially suitable for processing large-scale datasets. In the financial field, Naive Bayes algorithm is often used to solve various data classification problems, such as filtering high-quality or low-quality customer information, personalized sentiment analysis, and topic classification.

### 2.5 Support Vector Machine (SVM)

Support Vector Machine is an algorithm used for classification and regression. It separates samples of different categories by mapping

the input samples into a high-dimensional space and finding an optimal hyperplane. Support vector machines have the advantages of strong generalization ability and are not easily affected by overfitting, but they may also be limited due to overly strict assumptions about data distribution. In the financial field, support vector machines are often used to solve various classification and regression problems, such as image recognition, text classification, and bioinformatics.

### 2.6 Regression Algorithm

Common regression algorithms include linear regression and logistic regression. Linear regression fits a simulated line segment based on historical data to predict the classification or approximate values of unknown data.

Linear regression assumes a linear relationship between the dependent variable and the independent variable (Y=KX+B). The goal of linear regression is to find a line that minimizes the sum of distances from all data points to that line. If there is collinearity between the high-dimensional dataset and the independent variable, linear regression is transformed into ridge regression. Linear regression is often used for data prediction and classification.

The idea of logistic regression algorithm is similar, but it uses more complex S-shaped curves to fit the data, mapping the results of linear functions to Sigmoid functions. The Sigmoid function is often used as an activation function for neural networks, where the range of variable mappings is between (0, 1), representing the probability of occurrence. Logistic regression algorithm is an algorithm used for binary classification problems. It applies the sigmoid function to the output of linear regression, converts the output of linear regression into a probability value, and classifies and predicts the input samples based on this probability value.

The logistic regression algorithm has the advantages of simplicity, ease of implementation, and efficiency, and can also handle multi classification problems. Logistic regression algorithm is often used to solve binary classification problems in the financial field, such as data filtering, market forecasting, and product recommendation.

### 2.7 Neural Network Algorithm

Neural network algorithm is a computational model that simulates the neural network structure of the human brain, capable of simulating the memory and reasoning processes of the human brain. It classifies or regresses input samples by constructing a network composed of multiple interconnected neurons. Neural network algorithms have strong nonlinear mapping and generalization abilities, but they also face problems such as overfitting and gradient vanishing. In the financial field, neural network algorithms are often used to solve various classification and regression problems, such as image recognition, speech recognition, and natural language processing.

## 3. Advanced application skills of algorithms in production environments

### 3.1 Feature engineering

In production practice, feature engineering technology is particularly important. Advanced feature engineering techniques include a series of feature selection and transformation techniques aimed at improving model performance. Among them, feature selection methods such as filtering out redundant features, handling missing and outlier values, and feature encoding and transformation techniques are common operations.

In terms of feature selection, common algorithms include filtering, packaging, and embedding. The filtering algorithm selects features based on their statistical properties, such as correlation, variance, and mutual information. The wrapper algorithm uses supervised learning algorithms to score and select features. Embedded algorithms integrate the feature selection process into the model training process.

In terms of feature transformation, some common techniques include feature scaling, standardization, and normalization. Feature scaling can scale feature values to the same scale to avoid the influence of scale and dimension on the model. Standardization converts the eigenvalues into a standard normal distribution, so that each feature has the same weight. Normalization reduces the feature values to the range of [0, 1], making the algorithm more focused on local features.

### 3.2 Model Fusion and Stacking

Model fusion and stacking are commonly used techniques to improve model performance. Model fusion improves the expressive and generalization abilities of multiple models by fusing them together. Common fusion methods include series fusion, parallel fusion, and multi-level fusion. Tandem fusion concatenates multiple models together, with each model processing the input data once. Parallel fusion involves connecting multiple models in parallel, with each model processing the input data in parallel. Multi level fusion combines multiple models at multiple levels to adapt to different levels of data features.

Stacking is an ensemble learning technique that combines multiple learners to improve the generalization ability of a model. In stacking, low-level learners are responsible for learning the underlying features of data, while high-level learners learn based on the output of low-level learners.

### 3.3 Hyper parameter optimization techniques

Hyper parameter optimization is a key step in machine learning model training, aimed at finding the optimal hyper parameter configuration to improve model performance. Hyper parameters are parameters that need to be manually set during model training, such as learning rate, iteration count, and regularization strength.

Hyper parameter optimization methods include grid search, random search, and Bayesian optimization. Grid search searches for the optimal hyper parameter configuration by traversing a given range of hyper parameter combinations. Random search randomly samples points in the hyper parameter space and selects the optimal hyper parameter configuration. Bayesian optimization is based on Bayesian statistical theory, which iteratively updates the range of hyper parameters to find the optimal hyper parameter configuration.

In practice, the use of automated hyper parameter optimization techniques can significantly improve model performance. Common automated hyper parameter optimization tools include Hyperopt, TPOT, and AutoML.

### 3.4 Gradient Boosting Decision Tree (XGBoost) Application Techniques

XGBoost is an efficient gradient boosting decision tree algorithm with strong generalization ability and fast running speed. To improve the performance of XGBoost, the following points should be noted:

① Data preprocessing: Use appropriate data preprocessing methods such as feature scaling, discretization, and one hot encoding to meet the requirements of XGBoost algorithm.

② Adjusting parameters: XGBoost has many adjustable parameters, such as learning rate, maximum depth, regularization parameters, etc. Adjusting these parameters based on different datasets and task types can improve the performance of the model.

③ Multi task learning: XGBoost supports multi task learning and can process multiple tasks in the same model at the same time.

This article analyzes common machine learning algorithms in the financial field, which may not cover all but are representative. It should be noted that the future development direction of artificial intelligence in the financial field is mainly in the following aspects:

Efficient financial services: Artificial intelligence can help financial institutions provide financial services more efficiently, such as predicting customer needs through machine learning algorithms and personalized recommendation of financial products through deep learning algorithms.

Intelligent investment strategy: A recommendation system based on artificial intelligence can help investors formulate investment strategies. For example, analyzing market data through machine learning algorithms and predicting market trends through deep learning algorithms.

Financial System Security: AI can assist financial institutions in conducting big data analysis, predicting and preventing risks, and building a more secure financial system. For example, detecting financial fraud through machine learning algorithms and identifying financial risks through deep learning algorithms.

## 4. Conclusion

This article summarizes and explains the application methods of common machine learning algorithms in the financial field.The combination of artificial intelligence and financial analysis is reshaping the future of the financial industry. By improving efficiency, enhancing accuracy, and real-time analysis, AI provides strong support for investors, which cannot be separated from the support of algorithms. Looking ahead, artificial intelligence algorithms will play an increasingly important role in financial analysis, driving innovation and development in the financial industry.

## References

[1] Siddique, Nazmul, and H. Adeli. "Applications of Harmony Search Algorithms in Engineering." International Journal of Artificial Intelligence Tools 24.06(2015):1530002.

[2] Wang Hui and Ji Xiaoliang Analysis of Artificial Intelligence Classic Algorithms on Iris Dataset. "Microcomputer Information 000.018 (2021): 14-19, 21

[3] B, Rodolfo C. Cavalcante A, et al. "Computational Intelligence and Financial Markets: A Survey and Future Directions." Expert Systems with Applications 55.C(2016):194-211.

**Author Introduction:** Xiaoliang Ji, Senior Engineer, pursuing DBA at ENAE Business School. Research focuses on Big Data Analysis and Statistics.