Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.32885

Prediction of Low-Altitude Concept Company Characteristics Based on Machine Learning Algorithms

Yihang Yuan

International College, Zhengzhou University, Zhengzhou, China
18237113766@163.com

Abstract. The low-altitude economy, as an emerging industry form covering multiple fields such as aviation manufacturing, logistics transportation, and urban services, has become an important engine driving regional economic upgrading and industrial structure optimization. Accurately identifying low-altitude concept listed companies is an important foundation for conducting industry trend analysis, enterprise value assessment, and precise policy support. The machine learning classification algorithm, with its outstanding data mining and pattern recognition capabilities, has been widely applied in areas such as enterprise attribute prediction and industry classification, providing strong technical support for the identification of low-altitude concept listed companies. In response to the problem of low classification accuracy of low-altitude concept listed companies under high-dimensional imbalanced data, this paper proposes the KPCA-ISSA-RF classification algorithm. First, it conducts correlation analysis and violin plot analysis, and then uses multiple machine learning algorithms for comparative research. Experimental results show that the KPCA-ISSA-RF algorithm proposed in this paper has significantly better comprehensive performance than the AdaBoost, GBDT, decision tree, CatBoost, random forest, ExtraTrees, XGBoost, KNN, and logistic regression algorithms. Its accuracy and recall rate reach 92.9%, and the precision and F1 value are both 91.6%, ranking first among all algorithms. It demonstrates strong classification discrimination ability and comprehensive fitting effect, providing reliable technical support for related research and practice in the low-altitude economy.

Keywords: Low-altitude concept company characteristics, KPCA, ISSA.

1. Introduction

Low-altitude economy, as an emerging industry form integrating aviation manufacturing, logistics transportation, and urban services, has become an important engine driving regional economic upgrading and industrial structure optimization. Accurately identifying low-altitude concept listed companies is a core prerequisite for conducting industry trend analysis, enterprise value assessment, and precise policy support [1]. With the continuous growth of the number of listed companies in the capital market, the disclosed enterprise data from 2000 to 2024 cover multiple dimensions such as industry attributes, financial operations, and geographical distribution. Such data present characteristics of high feature dimensions and uneven category distribution, and contain a large

DOI: 10.54254/2755-2721/2026.32885

amount of nonlinear correlation information. Traditional enterprise classification methods mostly rely on manual feature selection, which is difficult to effectively handle redundant information in high-dimensional data and lacks targeted solutions for classification bias caused by data imbalance. This makes it unable to meet the actual needs of precise identification of low-altitude concept listed companies and urgently requires the construction of an efficient classification model adapted to high-dimensional imbalanced data [2].

Machine learning classification algorithms, with their powerful data mining and pattern recognition capabilities, have been widely applied in enterprise attribute prediction and industry classification, providing technical support for the identification of low-altitude concept listed companies [3]. Such algorithms can automatically learn the mapping relationship between features and target variables from massive data. The random forest algorithm, due to its advantages in handling nonlinear data, resisting overfitting, and outputting feature importance, has become a commonly used model for enterprise classification tasks [4]. However, when dealing with high-dimensional data including financial indicators and geographical features, directly using the random forest algorithm is prone to increase model computational complexity due to increased feature redundancy, resulting in decreased classification efficiency; at the same time, traditional parameter optimization methods such as grid search and genetic algorithms are prone to fall into local optimal solutions, unable to fully exert the model's performance, restricting the improvement of classification accuracy and generalization ability. It is necessary to improve the existing algorithms from the aspects of feature dimension reduction and parameter optimization [5].

To address the problem of insufficient classification accuracy of low-altitude concept listed companies under high-dimensional imbalanced data, this paper proposes the KPCA-ISSA-RF classification algorithm. This algorithm first uses kernel principal component analysis to reduce the original high-dimensional features, and through kernel function mapping, converts the linearly non-separable high-dimensional data into linearly separable data in the low-dimensional space, effectively eliminating feature redundancy and retaining key information; then, an improved sparrow search algorithm is introduced to optimize the core parameters of the random forest, such as the number of decision trees and node splitting thresholds, using the stronger global search ability and convergence speed of the improved algorithm to avoid the defect of traditional optimization methods falling into local optimal solutions; finally, through the collaborative effect of KPCA dimension reduction and ISSA parameter optimization, the accuracy and stability of the random forest in the classification task of low-altitude concept listed companies are improved, providing a more reliable technical tool for enterprise identification and industry analysis in the low-altitude economy field.

2. Data source and analysis

This dataset is constructed based on the low-altitude concept-related enterprise data collected from the public listed company information database from 2000 to 2024. During the data processing, the original information was first cleaned and integrated to eliminate invalid records. Then, a multi-dimensional feature system was constructed, which includes ID features (year, stock code, stock abbreviation, year of initial listing), company basic features (company listing years, number of employees, company size), industry and regional features (industry name, industry code, industry category, province, city, district and county, whether it is an eastern province, whether it is an economically developed region), and financial and operational features (era, whether it is recent, revenue growth rate, net profit rate, asset-liability ratio, R&D investment ratio, province code, city code, district and county code). The target variable is "whether it is a low-altitude concept listed

DOI: 10.54254/2755-2721/2026.32885

company in 2025", with a total of 3999 pieces of data. Among them, 3672 pieces (accounting for 91.82%) are non-low-altitude concept listed companies, and 327 pieces (accounting for 8.18%) are low-altitude concept listed companies. Firstly, correlation analysis was conducted for each variable, as shown in Figure 1.

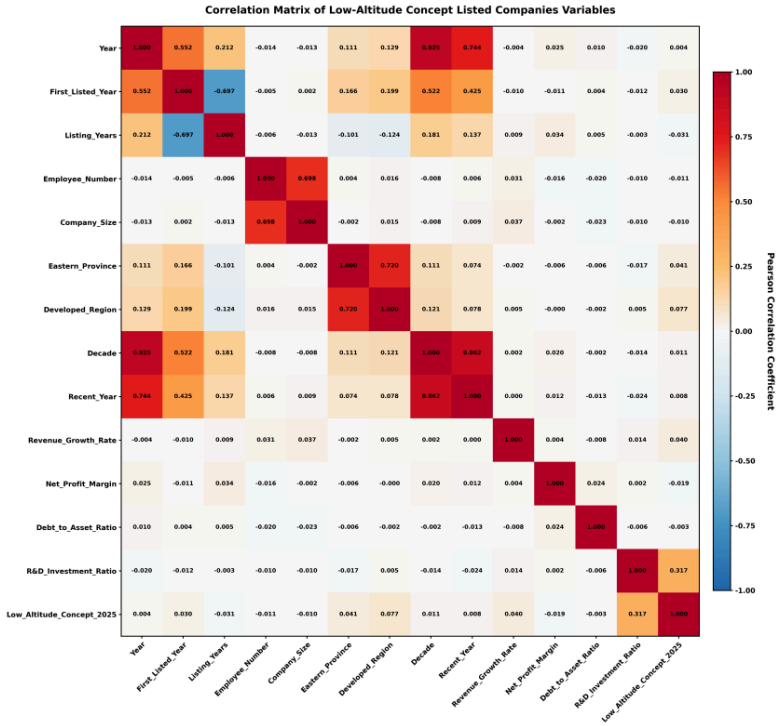

Figure 1. Correlation analysis

Perform violin analysis on each variable and draw violin plots, as shown in Figure 2.

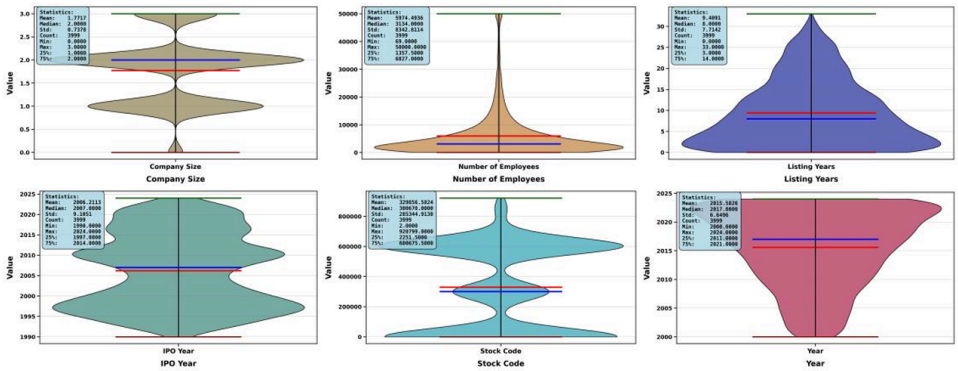

Figure 2. Violin plots

3. Method

3.1. KPCA

KPCA, or Kernel Principal Component Analysis, is a nonlinear extension of Principal Component Analysis in the kernel space. Traditional PCA projects high-dimensional data into a low-dimensional

DOI: 10.54254/2755-2721/2026.32885

subspace through linear transformation to extract the main features, but its ability to reduce dimensions is limited when dealing with nonlinear data [6]. KPCA avoids the complex calculations required for directly processing high-dimensional mappings by using kernel techniques. It implicitly maps the original data to a high-dimensional feature space through the kernel function and performs linear PCA processing in this space [7]. It first calculates the kernel function values between all data points to construct a symmetric kernel matrix, centers the kernel matrix, and then calculates the eigenvalues and eigenvectors. It selects the eigenvectors corresponding to the larger eigenvalues as the kernel principal components. Finally, it projects the original data onto these principal components to achieve nonlinear dimensionality reduction and effectively extract the nonlinear features of the original data.

3.2. ISSA

ISSA, which is the improved sparrow search algorithm, is an optimization improvement of the standard sparrow search algorithm. The standard SSA simulates the foraging behavior and avoidance of predators of sparrows, and completes optimization by having discoverers explore food sources, joiners follow the foragers, and sentinels monitor dangers. However, it has problems such as uneven initial population distribution, easy to fall into local optimum, and slower convergence speed [8]. ISSA optimizes through multiple strategies. Common improvements include using chaotic mapping and elite reverse learning to optimize the initial population distribution, introducing an adaptive discovery ratio that dynamically adjusts the number as the iteration progresses, designing dynamic weight factors to balance global exploration and local exploitation, combining Cauchy mutation or sine perturbation to break local optimum, and also adding the previous generation's global optimal solution to enhance the sufficiency of the search.

3.3. Random Forest

RF stands for Random Forest, which is a classic combination of Bagging integration and decision trees. Its core idea is that multiple decision trees work collaboratively. By randomly sampling both the samples and the features, multiple independent decision trees are constructed. Finally, the classification result is obtained through voting [9]. During the construction process, the original dataset is sampled with replacement to obtain multiple sub-sample sets. Each tree is trained using one of the sub-sample sets, and during each node split, the optimal partitioning feature is selected only from the randomly selected feature subset. After all the trees are trained, a forest is formed. During prediction, the new sample is input into each tree to obtain individual results, and the category with the highest occurrence frequency among all the results is taken as the final classification result [10]. The network structure of the Random Forest algorithm is shown in Figure 3.

DOI: 10.54254/2755-2721/2026.32885

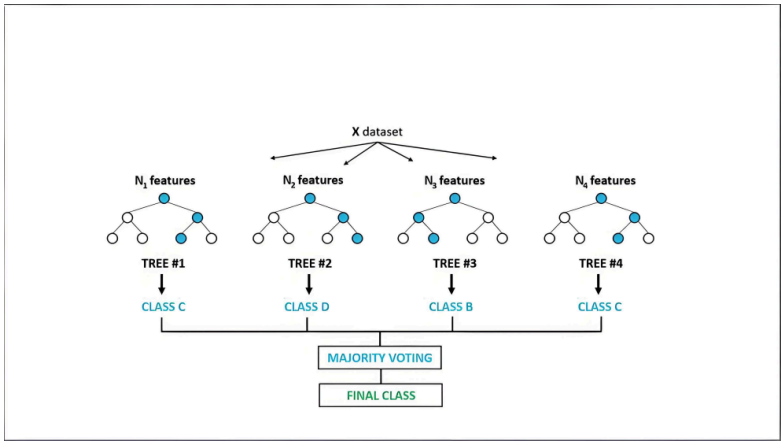

Figure 3. The network structure of the Random Forest algorithm

3.4. KPCA-ISSA-RF classification algorithm

KPCA-ISSA-RF is a hybrid classification algorithm that integrates kernel principal component analysis, improved sparrow search algorithm, and random forest. It accomplishes the classification task through three steps of collaboration. Firstly, KPCA is used to perform nonlinear dimensionality reduction on the original high-dimensional data, eliminating redundant information and noise, and retaining the most discriminative core features, thereby reducing the complexity and computational cost of subsequent model training. Then, ISSA is used to optimize the key parameters of the random forest. ISSA efficiently searches in the parameter space by simulating the improved behavior of the sparrow flock, finding the parameter combination that maximizes the classification accuracy of the random forest, solving the problem of random forest parameter setting being dependent on experience. Finally, the features after dimensionality reduction are input into the optimized random forest model by ISSA for training and classification. Combining the feature extraction ability of KPCA, the optimization ability of ISSA, and the strong classification ability of RF, the accuracy and stability of complex data classification tasks are significantly improved.

4. Result

A comparative experiment was conducted using multiple machine learning algorithms, and the performance indicators of each algorithm were presented, as shown in Table 1.

Table 1. The performance indicators of each algorithm

Model	Accuracy	Recall	Precision	F1	AUC
AdaBoost	0.921	0.921	0.901	0.905	0.967
GBDT	0.915	0.915	0.889	0.895	0.96
Decision Tree	0.9	0.9	0.889	0.894	0.892
CatBoost	0.917	0.917	0.902	0.885	0.963
Random Forest	0.912	0.912	0.92	0.872	0.962
ExtraTrees	0.917	0.917	0.903	0.88	0.966
XGBoost	0.919	0.919	0.904	0.898	0.965
KNN	0.912	0.912	0.869	0.873	0.931

DOI: 10.54254/2755-2721/2026.32885

Table 1. (continued)

Logistic Regression	0.912	0.912	0.833	0.871	0.929
KPCA-ISSA-RF	0.929	0.929	0.916	0.916	0.966

A bar chart presenting the comparison results of each indicator is shown in Figure 4.

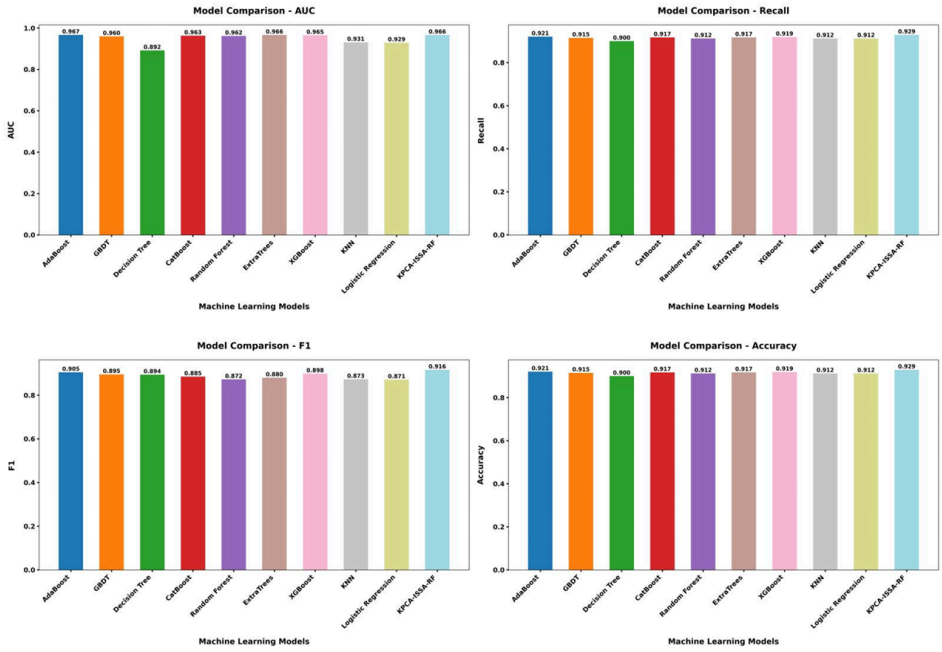

Figure 4. The bar chart presenting the comparison results of each indicator

Based on the experimental results of various machine learning algorithms, the KPCA-ISSA-RF algorithm proposed in this paper significantly outperforms the comparison algorithms such as AdaBoost, GBDT, decision tree, CatBoost, random forest, ExtraTrees, XGBoost, KNN, and logistic regression in terms of comprehensive performance. The accuracy rate of this algorithm reaches 92.9%, and the recall rate is also 92.9%, both of which are the highest among all algorithms. The precision rate is 91.6% and the F1 value is 91.6%, which are also far superior to other comparison algorithms. It demonstrates strong classification discrimination ability and comprehensive fitting effect, with an AUC value of 96.6%. It is in the same high-level range as the excellent AdaBoost and ExtraTrees algorithms. Compared with traditional decision tree algorithms, this algorithm has significantly improved in accuracy, recall rate, F1 value, and AUC value. Compared with classic classification algorithms such as KNN and logistic regression, as well as mainstream boosting tree algorithms such as GBDT and XGBoost, KPCA-ISSA-RF has achieved leading performance in various core evaluation indicators, fully verifying the effectiveness and superiority of this improved algorithm.

5. Conclusion

The low-altitude economy, as an emerging industry integrating aviation manufacturing, logistics transportation, and urban services, has become an important engine driving regional economic upgrading and industrial structure optimization. Accurately identifying companies related to the

DOI: 10.54254/2755-2721/2026.32885

low-altitude concept is the core prerequisite for conducting industry trend analysis, enterprise value assessment, and precise policy support.

Machine learning classification algorithms, with their powerful data mining and pattern recognition capabilities, are widely applied in areas such as enterprise attribute prediction and industry classification. They provide solid technical support for the identification of companies related to the low-altitude concept. In response to the problem of insufficient classification accuracy for such enterprises in high-dimensional imbalanced data, this paper proposes the KPCA-ISSA-RF classification algorithm. Through correlation analysis, violin plot analysis, and comparison of multiple machine learning algorithms, this algorithm demonstrates significantly superior comprehensive performance compared to algorithms such as AdaBoost and GBDT. Its accuracy rate and recall rate reach 92.9%, and the precision and F1 value are both 91.6%, showcasing strong classification discrimination and comprehensive fitting capabilities.

This algorithm effectively solves the classification problems brought about by high-dimensional imbalanced data, improving the accuracy of identifying companies related to the low-altitude concept. It provides more reliable technical support for industry development analysis and enterprise value assessment, and has significant practical significance for promoting the standardized development of the low-altitude economy industry and facilitating the precise implementation of policies.

References

[1] Sun, Xiting, et al. "Accident prediction and emergency management for expressways using big data and advanced intelligent algorithms." 2025 IEEE 3rd International Conference on Image Processing and Computer Applications (ICIPCA). IEEE, 2025.

[2] Sayarshad, Hamid R. "Designing an intelligent emergency response system to minimize the impacts of traffic incidents: a new approximation queuing model." International journal of urban sciences 26.4 (2022): 691-709.

[3] Grigorev, Artur, Adriana-Simona Mihaita, and Fang Chen. "Traffic Incident Duration Prediction: A Systematic Review of Techniques." Journal of Advanced Transportation 2024.1 (2024): 3748345.

[4] Yang, Samgyu. "Developing Real-Time Crash Prediction System using AI-Based Methods on Interstate Incorporating Express Lanes." (2025).

[5] Abraham, Anuj, Yi Zhang, and Shitala Prasad. "Evacuation management framework towards smart city-wide intelligent emergency interactive response system." arXiv preprint arXiv: 2403.07003 (2024).

[6] Grigorev, Artur, et al. "Automatic accident detection, segmentation and duration prediction using machine learning." IEEE Transactions on Intelligent Transportation Systems 25.2 (2023): 1547-1568.

[7] Yemelyanenko, Sergiy, et al. "Improving the operational efficiency of control centers for emergency events by using GIS technologies." (2023).

[8] Peelam, Mritunjay Shall, et al. "A review on emergency vehicle management for intelligent transportation systems." IEEE transactions on intelligent transportation systems (2024).

[9] Adewopo, Victor A., and Nelly Elsayed. "Smart city transportation: Deep learning ensemble approach for traffic accident detection." IEEE Access 12 (2024): 59134-59147.

[10] Al-Zabidi, Ayoub, et al. "Statistical modeling of emergency medical services' response and rescue times to road traffic crashes in the Kingdom of Saudi Arabia." Case Studies on Transport Policy 10.4 (2022): 2563-2575.