Scientific
Research
Publishing

A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method

Yan Zhang[1], Lin Chen[2*]

[1]Investment Bank Department, Chengdu Agricultural and Commercial Bank, Chengdu, China
[2]School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China
Email: *Chenlin2@uestc.edu.cn

Abstract

China's bond market is an emerging market. The number of bond defaults has been increasing in recent years, but the data set is severely imbalanced. Based on financial data of total 6731 corporate bond issuers which 50 bond issuers had defaulted, this paper uses the XGboost algorithm and an Over-sampling method named SMOTE to predict the default of bond issuers. The results show that the XGboost algorithm has advantages over the traditional algorithm in processing imbalanced data, and SMOTE is one of the effective methods to deal with imbalanced samples. Then, this is an effective way to predict the default risk of bond issuers in an emerging market.

Keywords

XGboost, Default Risk, Bond Issuer, Imbalanced Data, SMOTE

1. Introduction

Corporate credit risk is one of the most key risks for financial institutions and investors. Besides two key theoretical models for evaluating credit risk, one of which is called structural mode (Black and Scholes, 1973; Merton, 1974; Black and Cox, 1976) and the other is called reduced model (Jarrow and Turnbull, 1995; Longstaff and Schwartz, 1995; Jarrow, Lando, and Turnbull, 1997; Duffee, 1999; Duffie and Singleton, 1999; Jarrow and Turnbull, 2000; Duffie and Lando, 2001). A lot of statistical models have been used to predict the corporate credit risk (Altman, 1968; Altman, Haldeman, and Narayanan, 1977; Martin, 1977; Ohlson, 1980; Kim and Expert, 1999; Sung, Chang, and Lee, 1999; Shah and Murtaza, 2000; Nanda and Pendharkar, 2001).

In recent years, a machine learning method named eXtreme Gradient Boosting (XGBoost) has been used to evaluate personal credit risk (Li et al., 2020; Ma et al., 2018; Xia, Liu, and Liu, 2017; Chang, Chang, and Wu, 2018). Since Chen proposed the eXtreme Gradient Boosting (XGBoost) algorithm in 2016 (Chen and Guestrin, 2016), it has been one of the most important, effective, and widely used machine learning methods. The XGBoost has been used in many fields such as disease diagnosis (Liu et al., 2021; Zhang et al. 2020b; Zhang, Deng, and Jia, 2020a; Zhang and Gong, 2020), image recognition (Zhao, Guo et al., 2020; Zhao, Zeng et al., 2020), traffic flow prediction (Zhang and Zhang, 2020), sports analytics (Yigit, Samak, and Kaya, 2020), power system (Xue and Wu, 2020; Raichura, Chothani, and Patel, 2021; Li et al., 2018), e-commerce (Song and Liu, 2020), personal credit (Li et al., 2020; Ma et al., 2018; Xia, Liu, and Liu, 2017; Chang, Chang, and Wu, 2018) forecasting of weather (Jin et al., 2020; Fan et al., 2020; Zheng and Wu, 2019), public security (Feng et al., 2020), fault detection (Zhang et al., 2018; Chen et al., 2019; Lei et al., 2019; Lin et al., 2019) and so on. Those studies show that XGBoost can improve the prediction accuracy better than other statistical methods.

As of December 31, 2020, there are 57,039 issued bonds in the China bond market, and the value of bonds has reached 114.33 trillion yuan. Since 2014, the number of default bonds has increased year by year. There were 145 default bonds in 2020, and the face values of default bonds were 164.7 billion. The default in the bond market caused huge losses to investors, and there are new features of default bonds; for example, many local and central state-owned corporates' bonds default and high-rating bonds default. So, this paper aims to use XGBoost to forecast the default of bond issuer in China. Moreover, the number of default bonds is still small relative to all samples, the imbalanced data classification is the biggest challenge for modeling. Many Resampling techniques have been developed in the past two decades to cope with imbalanced data classification. Those techniques fall into three groups: Over-sampling methods, Under-sampling methods and Hybrid methods (Chawla et al., 2002; Chawla et al., 2003; Guo et al., 2017). This paper will apply the Over-sampling methods to cope with imbalanced data.

The rest of this paper is organized as follows. Section 2 describes the basic principle of XGBoost and SMOTE. Section 3 describes the data set. Section 4 describes how to use XGBoost to predict defaulted bonds issuer. Section 5 is the conclusion.

2. XGboost Algorithm and SMOTE Algorithm

The XGboost algorithm adopts the idea of integration, and it can be used to solve both classification and regression problems. The algorithm mainly uses the integration idea to solve the minimum loss function through the second-order Taylor expansion, determines the split node, and constructs the final model. The basic algorithm as follows (Chen and Guestrin, 2016):

$$\hat{y}_i = \sum_k^K f_k(x_i) \tag{1}$$

In Equation (1), f_k is an additive function and represents a tree, then it is a tree ensemble model and uses K additive functions to predict the output.

To learn the set of functions used in the model, the model minimizes the following regularized objective.

$$L(\varnothing) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{2}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\gamma \|\omega\|^2$$

In Equation (2), $l(y_i, \hat{y}_i)$ is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and target y_i, and the $\Omega(f)$ penalizes the model's complexity. The additional regularization term helps to smooth the final learnt weights to avoid over-fitting. T represents the number of nodes on the tree, and ω represents nodes' values on the tree.

Equation (2) includes functions as parameters and cannot be optimized using traditional Euclidean space optimization methods. Chen & Guestrin (2016) put forward a formula for evaluating the split candidates of the tree structure.

In short, Although XGboost is developed from gradient tree boosting algorithms, XGboost is still better than traditional gradient tree boosting algorithms for the following reasons: 1) The regular term has been added to the cost function of XGboost to control the complexity of the model, and make the learned model simpler, and prevent overfitting. 2) The base learning instrument of XGboost can be both linear classification and CATR classification. The traditional gradient tree boosting algorithms only uses CATR classification. 3) XGboost learns from random forests and supports column sampling, which reduces overfitting and reduces computation.

SMOTE (Synthetic minority over-sampling technique) is one of Over-sampling methods that a new minority class sample is created in the neighborhood of the minority class sample under consideration (Chawla et al., 2002). Using this technique, a new minority class sample is created in the neighborhood of the minority class sample under consideration. The algorithm of synthesis is to randomly select a minority sample A and its nearest neighbor B, between A and B as a newly synthesized minority sample. For example, the new minority sample C is synthetized as:

$$C = A + rand(0,1) \times |A - B| \tag{3}$$

In Equation (3), $rand(0,1)$ denotes a random number between 0 and 1, $|A - B|$ denotes the Euclidean distance for the continuous features.

Now XGboost algorithm and SMOTE algorithm can be implemented using python packages. SMOTE algorithm is used to deal imbalanced data sets, and the XGboost algorithm is used to classify the samples.

3. Data Set

We have got financial data of total 6731 bond issuers from the wind database in

2018, which has 50 bond issuers had defaulted in 2019. The data is typical imbalanced data; we will use SMOTE to deal with this imbalanced data. According to existing research, finally this paper selected 16 financial indicators based on analysis of four aspects of bond issuer from profitability, operational capacity, solvency capacity, capital structure (Table 1).

There are some key indicators that have a great effect on the accuracy of prediction: 1) Monetary/total debt is a measure of the solvency of the issuing company. The ratio can reflect the degree of protection of the company's cash and cash equivalents to the company's current liabilities. It is generally believed that the larger the ratio, the stronger the solvency of the company. 2) Return on asset and return on equity are indicators to measure the profitability of issuing companies relative to their total assets or equity. Generally speaking, the stronger the profitability of the company, the higher the ability to repay debts, and the smaller the risk of default. 3) The rate of stock turnover indicates how fast a company sells inventory. The ratio reflects the operational capacity of issuing companies. In general, the stronger the operational capacity of the company, the higher the ability to repay debts, and the smaller the risk of default. 4) The asset-liability ratio is a measure of capital structure.It is a leverage ratio that defines the total amount of debt relative to assets owned by a company. The lower the asset-liability ratio, the more repayment ability of the company and the less probability to default.

Table 1. The statistical description of financial data 6731 bond issuers (%).

Financial indicators \ Statistical description	min	max	average	variance
profit rate of core businesses	−25,406.88	15,244.17	11.20	260,017.10
income growth rate of core businesses	−181.99	346,306.17	172.91	34,596,379.30
return on asset	−109.20	48.19	3.09	35.04
return on equity	−832.57	138.85	3.68	463.02
EBITDA/Total Income	−23,571.09	16,280.29	26.27	205,915.70
operating cash/EBITDA	−12,286.60	2680.67	−1.80	30,347.40
liquidity ratio	0.05	5009.00	5.37	7331.34
quick ratio	0.02	5009.00	3.86	7266.21
rate of stock turnover	−16.11	48,228.55	37.28	524,720.59
asset-liability ratio	0.02	800.92	58.81	478.51
short-term debt/total debt	0.10	100.00	56.71	621.07
debt with interest/total investment capital	0.00	3308.84	41.66	2415.57
monetary/total debt	0.00	5009.00	1.99	6575.71
interest multiples	−2357.01	35,805.41	61.86	857,667.92
EBITDA/total debt with interest	−12,218.75	378,162.66	172.32	29,637,575.78

4. Effectiveness Analysis of XGboost Algorithm

The classification model performance could be evaluated by some indicators such as Precision, Accuracy, Recall, F-Measure. This paper mainly uses the area under the curve of ROC (Receiving Operating Characteristic) as the performance evaluation of the classification model.

Suppose the issuer of the defaulted bond is a positive sample (positives), and the issuer without defaulted bond is a negative sample (negatives). TP indicates the number of samples that the issuer of a defaulted bond is correctly predicted. FN indicates the number of samples that the issuer of a defaulted bond is not correctly predicted. FP indicates the number of samples that the issuer without defaulted bond is not correctly predicted. TN indicates the number of samples that the issuer without defaulted bond was correctly predicted.

True Positive Rate (TPR) is defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

False Positive Rate (FPR) is defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

For every classification model, the ROC curve is formed by a pair of FPR and TPR. ROC curve is a commonly used dichotomous model evaluation standard, which is used to show whether the effect is good or not. Since the ROC curve cannot quantitatively evaluate the classifier, the effect of the model is generally measured by the AUC (Area under the ROC curve), and the AUC value can well describe the overall performance of the model. The larger the FPR, the more real negative classes in the predicted positive class, and the larger the TPR, the more real positive classes in the predicted positive class. The closer the ROC curve is to the upper left corner, the greater the AUC value, and the model is better.

The following algorithm is implemented based on the Python software package. Total samples were classified according to 20% of the test samples and 80% of the training samples. There are six important parameters[1] for tree Booster. 1) Learning rate that step size shrinkage used in update to prevents overfitting, this parameter is 0.05; 2) Maximum depth of a tree, this parameter is 5; 3) Minimum sum of instance weight (hessian) needed in a child, this parameter is 3; 4) Minimum loss reduction required to make a further partition on a leaf node of the tree, this parameter is 0.8; 5) Maximum delta step we allow each leaf output to be, this parameter is 0 ; 6) Subsample ratio of the training instances,this parameter is 1. Those parameters are settled by grid search algorithm.

At first, without the imbalanced treatment of the data, the ROC curves and AUC are obtained in Figure 1.

Then before using the XGboost algorithm, we first use SMOTE to rebalance training samples. After that, the results on the test set are shown in Figure 2.

---

[1]https://xgboost.readthedocs.io/en/latest/parameter.html.

Finally, we compare the XGboost with other classification methods such as the K-Nearest Neighbors (KNN), Logistic regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and Random Forest (RF). The AUC of the six modes list in the Table 2.

According to Table 2, even on very imbalanced data sets, XGboost still gets better results than traditional classification algorithms. It is effective in improving the effect of classification algorithm by data rebalancing.
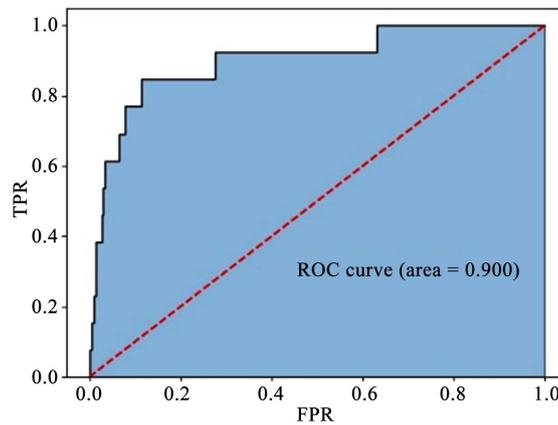

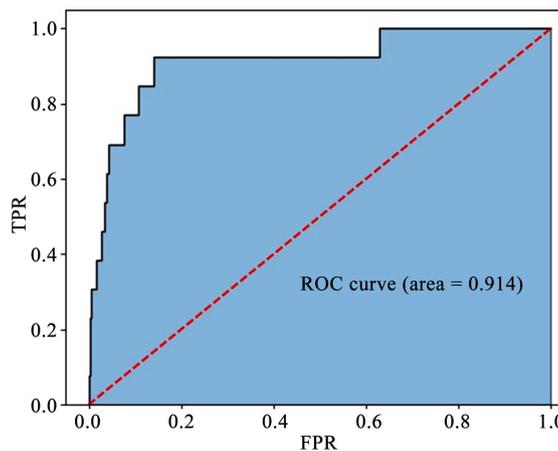**Figure 1.** The AUC without deal with imbalanced data.

Figure 2. The AUC with deal with imbalanced data.

Table 2. The AUC of different model.

Model	AUC of imbalanced data	AUC of balanced data
LR	0.733	0.858
SVM	0.682	0.694
KNN	0.722	0.782
DT	0.494	0.646
RF	0.774	0.836
XGboost	0.900	0.914

5. Conclusion

China's bond market is an emerging market. The time of development is short, so the default data show a severe imbalance feature. As the result of Section 4, some traditional statistical classification methods can't be used to classify the imbalanced data. This paper applies the XGboost algorithms to predict the default risk of corporate bond issuers in China. The conclusions are as follows. At first, the XGboost algorithm has advantages over the traditional algorithm in processing imbalanced data. Secondly, it is necessary to rebalance the imbalanced data by some method before building the prediction model. This paper shows SMOTE is an effective method to deal with imbalanced samples.

Funding

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance, 23,* 589-609.
https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations. *Journal of Banking and Finance, 1,* 29-54.
https://doi.org/10.1016/0378-4266(77)90017-6

Black, F., & Cox, J. C. (1976). Valuing Corporate Securities: Some Effects of Bond Indenture Provisions. *Journal of Finance, 31,* 351-367.
https://doi.org/10.1111/j.1540-6261.1976.tb01891.x

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy, 81,* 637-654. https://doi.org/10.1086/260062

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme Gradient Boosting Trees in the Construction of Credit Risk Assessment Models for Financial Institutions. *Applied Soft Computing, 73,* 914-920. https://doi.org/10.1016/j.asoc.2018.09.029

Chawla, N. V. et al. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research, 16,* 321-357. https://doi.org/10.1613/jair.953

Chawla, N. V. et al. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *Knowledge Discovery in Databases: Pkdd 2003, Proceedings,* Vol. 2838, 107-119. https://doi.org/10.1007/978-3-540-39804-2_12

Chen, M. H. et al. (2019). XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. *IEEE Access, 7,* 13149-13158. https://doi.org/10.1109/ACCESS.2019.2893448

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Francisco, 13-17 August 2016, 785-794.

https://doi.org/10.1145/2939672.2939785

Duffee, G. R. (1999). Estimating the Price of Default Risk. *The Review of Financial Studies, 12,* 197-226. https://doi.org/10.1093/rfs/12.1.197

Duffie, D., & Lando, D. (2001). Term Structures of Credit Spreads with Incomplete Accounting Information. *Econometrica, 69,* 633-664.
https://doi.org/10.1111/1468-0262.00208

Duffie, D., & Singleton, K. J. (1999). Modeling the Term Structure of Defaultable Bonds. *Review of Financial Studies, 12,* 687-720. https://doi.org/10.1093/rfs/12.4.687

Fan, Z. Y. et al. (2020). Estimating PM2.5 Concentrations Using Spatially Local Xgboost Based on Full-Covered SARA AOD at the Urban Scale. *Remote Sensing, 12,* 3368.
https://doi.org/10.3390/rs12203368

Feng, Y. et al. (2020). An XGBoost-Based Casualty Prediction Method for Terrorist Attacks. *Complex & Intelligent Systems, 6,* 721-740.
https://doi.org/10.1007/s40747-020-00173-0

Guo, H. X. et al. (2017). Learning from Class-Imbalanced Data: Review of Methods and Applications. *Expert Systems with Applications, 73,* 220-239.
https://doi.org/10.1016/j.eswa.2016.12.035

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing Derivatives on Financial Securities Subject to Credit Risk. *The Journal of Finance, 50,* 53-85.
https://doi.org/10.1111/j.1540-6261.1995.tb05167.x

Jarrow, R. A., & Turnbull, S. M. (2000). The Intersection of Markets and Credit Risk. *Journal of Banking & Finance, 24,* 271-299.
https://doi.org/10.1016/S0378-4266(99)00060-6

Jarrow, R. A., Lando, D., & Turnbull, S. M. (1997). A Markov Model for the Term Structure of Credit Risk Spreads. *The Review of Financial Studies, 10,* 481-523.
https://doi.org/10.1093/rfs/10.2.481

Jin, Q. W. et al. (2020). Estimating Tropical Cyclone Intensity in the South China Sea Using the XGBoost Model and FengYun Satellite Images. *Atmosphere, 11,* 423.
https://doi.org/10.3390/atmos11040423

Kim, C., & Expert, R. (1999). Linear Models and Nonlinear Models of Expert Decision Making in Bankruptcy Prediction: A Lens Models Analysis. *Journal of Management Information Systems, 16,* 189-206. https://doi.org/10.1080/07421222.1999.11518239

Lei, Y. F. et al. (2019). Fault Diagnosis Method for Hydraulic Directional Valves Integrating PCA and XGBoost. *Processes, 7,* 589. https://doi.org/10.3390/pr7090589

Li, C. B. et al. (2018). Predicting Short-Term Electricity Demand by Combining the Advantages of ARMA and XGBoost in Fog Computing Environment. *Wireless Communications & Mobile Computing, 2018,* Article ID: 5018053.
https://doi.org/10.1155/2018/5018053

Li, H. et al. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems, 35,* 52-61. https://doi.org/10.1109/MIS.2020.2972533

Lin, J. W. et al. (2019). Efficient Fault-Tolerant Routing in IoT Wireless Sensor Networks Based on Bipartite-Flow Graph Modeling. *IEEE Access, 7,* 14022-14034.
https://doi.org/10.1109/ACCESS.2019.2894002

Liu, P. et al. (2021). Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer. *IEEE Transactions on Biomedical Engineering, 68,* 148-160. https://doi.org/10.1109/TBME.2020.2993278

Longstaff, F. A., & Schwartz, E. S. (1995). A Simple Approach to Valuing Risky Fixed and

Floating Rate Debt. *Journal of Finance, 50,* 789-819. https://doi.org/10.1111/j.1540-6261.1995.tb04037.x

Ma, X. J. et al. (2018). Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electronic Commerce Research and Applications, 31,* 24-39. https://doi.org/10.1016/j.elerap.2018.08.002

Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression Approach. *Journal of Banking and Finance, 2,* 249-276. https://doi.org/10.1016/0378-4266(77)90022-X

Merton, R. C. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance, 29,* 449-470. https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

Nanda, S., & Pendharkar, P. (2001). Linear Models for Minimizing Misclassification Costs in Bankruptcy Prediction Finance and Management. *International Journal of Intelligent Systems in Accounting, 10,* 155-168. https://doi.org/10.1002/isaf.203

Ohlson, J. S. (1980). Financial Ratio and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, 19,* 109-131. https://doi.org/10.2307/2490395

Raichura, M., Chothani, N., & Patel, D. (2021). Efficient CNN-XGBoost Technique for Classification of Power Transformer Internal Faults against Various Abnormal Conditions. *IET Generation Transmission & Distribution, 15,* 972-985. https://doi.org/10.1049/gtd2.12073

Shah, J., & Murtaza, M. (2000). A Neural Network Based Clustering Procedure for Bankruptcy Prediction. *American Business Review, 18,* 80-86.

Song, P. Y., & Liu, Y. T. (2020). An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms. *Tehnicki Vjesnik-Technical Gazette, 27,* 1467-1471. https://doi.org/10.17559/TV-20200808113807

Sung, T., Chang, N., & Lee, G. (1999). Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction. *Journal of Management Information Systems, 16,* 63-85. https://doi.org/10.1080/07421222.1999.11518234

Xia, Y., Liu, C., & Liu, N. (2017). Cost-Sensitive Boosted Tree for Loan Evaluation in Peer-to-Peer Lending. *Electronic Commerce Research and Applications, 24,* 30-49. https://doi.org/10.1016/j.elerap.2017.06.004

Xue, W. L., & Wu, T. (2020). Active Learning-Based XGBoost for Cyber Physical System against Generic AC False Data Injection Attacks. *IEEE Access, 8,* 144575-144584. https://doi.org/10.1109/ACCESS.2020.3014644

Yigit, A. T., Samak, B., & Kaya, T. (2020). An XGBoost-Lasso Ensemble Modeling Approach to Football Player Value Assessment. *Journal of Intelligent & Fuzzy Systems, 39,* 6303-6314. https://doi.org/10.3233/JIFS-189098

Zhang, D. H. et al. (2018). A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost. *IEEE Access, 6,* 21020-21031. https://doi.org/10.1109/ACCESS.2018.2818678

Zhang, D. Y., & Gong, Y. C. (2020). The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. *IEEE Access, 8,* 220990-221003. https://doi.org/10.1109/ACCESS.2020.3042848

Zhang, X. J., & Zhang, Q. R. (2020). Short-Term Traffic Flow Prediction Based on LSTM-XGBoost Combination Model. *Cmes-Computer Modeling in Engineering & Sciences, 125,* 95-109. https://doi.org/10.32604/cmes.2020.011013

Zhang, X. X., Deng, T., & Jia, G. Z. (2020a). Nuclear Spin-Spin Coupling Constants Prediction Based on XGBoost and LightGBM Algorithms. *Molecular Physics, 118,* 1-10.

https://doi.org/10.1080/00268976.2019.1696478

Zhang, Y. L. et al. (2020b). A Novel XGBoost Method to Identify Cancer Tissue-of-Origin Based on Copy Number Variations. *Frontiers in Genetics, 11,* Article ID: 585029. https://doi.org/10.3389/fgene.2020.585029

Zhao, S. Q., Zeng, D. G. et al. (2020). Mutation Grey Wolf Elite PSO Balanced XGBoost for Radar Emitter Individual Identification Based on Measured Signals. *Measurement, 159,* Article ID: 107777. https://doi.org/10.1016/j.measurement.2020.107777

Zhao, W. Y., Guo, Y. X. et al. (2020). Fast Intelligent Cell Phenotyping for High-Throughput Optofluidic Time-Stretch Microscopy Based on the XGBoost Algorithm. *Journal of Biomedical Optics, 25,* 1-12. https://doi.org/10.1117/1.JBO.25.6.066001

Zheng, H., & Wu, Y. H. (2019). A XGBoost Model with Weather Similarity Analysis and Feature Engineering for Short-Term Wind Power Forecasting. *Applied Sciences-Basel, 9,* 3019. https://doi.org/10.3390/app9153019