Journal of Combinatorial Mathematics and Combinatorial Computing, 117: 159–168
DOI:10.61091/jcmcc117-14
http://www.combinatorialpress.com/jcmcc
Received 11 September 2023, Accepted 24 November 2023, Published 2 December 2023

Article

Credit rating algorithm of corporate bonds based on Gaussian process mixture model and improved K-means

Wenzhong Xia[1*]

[1] School of Zhangjiakou Vocational and Technical College, Zhangjiakou 075000, China.

* **Correspondence:** xwz8882001@163.com.

Abstract: The primary challenge in credit analysis revolves around uncovering the correlation between repayment terms and yield to maturity, constituting the interest rate term structure-an essential model for corporate credit term evaluation. Presently, interest rate term structures are predominantly examined through economic theoretical models and quantitative models. However, predicting treasury bond yields remains a challenging task for both approaches. Leveraging the clustering analysis algorithm theory and the attributes of an insurance company's customer database, this paper enhances the K-means clustering algorithm, specifically addressing the selection of initial cluster centers in extensive sample environments. Utilizing the robust data fitting and analytical capabilities of the Gaussian process mixture model, the study applies this methodology to model and forecast Treasury yields. Additionally, the research incorporates customer credit data from a property insurance company to investigate the application of clustering algorithms in the analysis of insurance customer credit.

Keywords: cluster analysis, k-means clustering algorithm, treasury bond yields, parameter learning, forecasting

1. Introduction

Credit represents a class of marketable securities [1], offering holders stable cash flow returns at specific future times [2]. Various factors influence bonds, encompassing both micro-entities and the macro-environment [3]. Credit categories include national bonds, policy bank financial bonds, corporate bonds, and municipal bonds based on the issuing entity [4]. Government bonds, issued based on the nation's credit, possess the highest credit rating [5]. Owing to their distinctive issuing entity, government bonds frequently serve as benchmarks for pricing other types of credit. The key determinants of credit value in practice include the issuing entity, denomination, coupon interest, repayment method, repayment period, and yield. Interest refers to the compensation received by the lender over a specific period, and the ratio of interest to the lent amount over that time is the interest rate or rate of return on funds [6].

Corporate credit rating is a management activity where an independent social intermediary assesses a company's borrowing and lending behavior's reliability and safety, providing an assessment report with professional symbols according to a specified methodology [7]. In essence, it is an evaluation of the enterprise's creditworthiness to repay principal and interest as promised, assessing the credit risk of the bond [8]. The credit rating solely judges the credit risk of the issued credit and does not reflect the rated credit's profitability and liquidity level. Therefore, rating results aid credit investors in gauging credit risk but should not be the sole basis for credit buying, selling, or holding decisions [9].

Given that ratings only assess a credit's risk without considering other factors like market price, supply and demand, and investor preferences, they serve as just one factor in investment decisions, not the sole basis [10]. When making credit investment decisions, investors must consider both the risk and return aspects of credit.

Credit ratings have a validity period, reflecting a specific credit's creditworthiness only during that period. Even within this timeframe, a credit's rating may change due to external environmental and internal operational conditions of the debt issuer [11].

A rating agency holds no legal responsibility for an investor's use of a rating. A credit rating from an agency serves as an indication to investors regarding the risk profile of various credits. It represents the agency's opinion, and investors are not obliged to share or adhere to it. Legally, there is no direct connection between the rating agency and the consequences investors face when using rating results [12].

The Gaussian process mixture model (MGP) is a potent statistical learning tool with robust learning and fitting capabilities. MGP models effectively describe multimodal data and reflect data volatility. They can be categorized into generative and discriminative models from the generative process perspective and into mixing in the time domain (MGP models) and mixing in the output space (mixGP models) from the mixing mode perspective.

The enhancement of the corporate credit rating system has spurred considerable scholarly interest in the rating methodology, a pivotal component of the system. Fitzpatrick (1932) conducted a univariate bankruptcy prediction study using ratios like net income to stockholders' equity and stockholders' equity to debt to predict firm bankruptcy [13]. Another study by [14] resulted in the well-known Z-score model and ZETA credit risk model, which utilized multivariate discriminant analysis for rating debt securities. Neural network analysis was applied to predict the financial crisis of Italian companies in [15]. In recent years, domestic scholars have delved deeper into this area, as seen in [16], which employed the internal rating method to enhance the current credit rating method of commercial banks in China [17]. This method considers not only the target data but also the relevance of each indicator, providing more valuable information and credit insights [18].

Given the limitations of financial factors in corporate credit rating analysis, such as lag, incompleteness (due to the largely incomplete or even false information disclosed in financial statements), and short-term focus, scholars are increasingly focusing on the role of non-financial factors in corporate credit rating. They argue that credit-issuing companies operate in an open system, subject to external factors, making non-financial factors early warning signs of future loan risk [19, 20].

This paper employs the MGP model to analyze corporate credit term structure data. Treasury yield data represent ”time-flow” data, with each data point correlated with neighboring points. This correlation is depicted by the covariance matrix of the MGP model. Due to policy influences and other factors at different time points, the volatility of Treasury yield data varies over time. The MGP model captures this differential volatility by expressing it through each GP component separately. These components describe local variations and are combined to enhance the MGP model's overall representation of data variability.

2. Gaussian Process Mixing Model

2.1. Gaussian Process (GP) Model

Mathematically, $Y(X)$ is considered a Gaussian process if, for any given N and $X = (x_1, \cdots, x_N)$, the corresponding $Y = (y_1, \cdots, y_N)$ follows a Gaussian distribution. In mathematical terms, a Gaussian process can be expressed as:

$$Y(X) \sim GP\left(m(X), K\left(X, X^{'}\right)\right). \tag{1}$$

In general problems, it is often assumed that $m(X) = 0$. In this paper, the Squared Exponential (SE) covariance function is utilized:

$$K\left(x, x^{'}\right) = \sigma_1^2 \exp\left(-\frac{\sigma_2^2}{2}\left\|x - x^{'}\right\|^2\right) + \sigma_3^2 I_{x=x'}. \tag{2}$$

For ease of representation, let the parameter be $\boldsymbol{\theta} = \left(\sigma_1^2, \sigma_2^2, \sigma_3^2\right)$, and the parameter learning of the GP model is efficiently performed using the maximum likelihood estimation algorithm.

2.2. MGP Model

In this paper, an MGP model in the form of a generative model is used, where each GP component is independent of each other. It is assumed that the Gaussian mixture model includes C Gaussian components, and each GP model is denoted as GPC. The Gaussian process mixture model generates the sample dataset $D = \{(x_n, y_n) \mid n = 1, \cdots N\}$ with the following rules:

1. First, the hidden variable z_n^c is introduced to describe the attribution of the sample to the GP component and follows the following distribution:

$$z_n^c = \begin{cases} 1; & (x_n, y_n) \in \mathrm{GP}_c \\ 0; & (x_n, y_n) \notin \mathrm{GP}_c \end{cases}, \tag{3}$$

   where $p\left(z_n^c = 1\right) = \pi_c, \sum_{c=1}^{C} \pi_c = 1$.
2. Under the condition $z_n^c = 1$, the sample input x_n follows a normal distribution with a mean of μ_c and a covariance of S_c^2.

$$P\left(x_n z_n^c = 1\right) \sim N\left(\mu_c, S_c^2\right). \tag{4}$$

3. Define $Z_c = \{n z_n^c = 1, n = 1, \cdots, N\}$, $X_c = \{x_n \mid Z_c\}$, $Y_c = \{y_n \mid Z_c\}$ as the sample label, input, and output of the c-th GP component, respectively. The c-th GP is defined as follows:

$$Y_c \sim GP\left(0, K\left(X_c, X_c \boldsymbol{\theta}_c\right)\right). \tag{5}$$

From the above three steps, we can see that the information flow direction of the MGP model is ”$Z \rightarrow X \rightarrow Y$”, which is consistent with the characteristics of the Treasury yield and the application scenario of the MGP model. Based on the dataset D, it is easy to derive the following log-likelihood function of MGP:

$$\log p(\boldsymbol{\Theta}, \boldsymbol{\Psi}\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \sum_c \left\{\sum_n \left[z_n^c \log\left(\pi_c p\left(x_n \mid \mu_c, S_c^2\right)\right)\right] + \log p\left(Y_c \mid X_c, \boldsymbol{\theta}_c\right)\right\}, \tag{6}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_c \mid c = 1, \cdots, C\}$ and $\Psi = \{\mu_c, S_c^2, \pi_c \mid c = 1\}$, denote the hyperparameters and parameters in the MGP model, respectively.

2.3. Algorithm Design

In this paper, we use the EM algorithm to learn hyperparameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_c \mid c = 1, \cdots, C\}$ and $\Psi = \{\mu_c, S_c^2, \pi_c \mid c = 1\}$. In practice, the main algorithms for learning parameters of MGP models are the MCMC algorithm, the variational Bayesian (VB) algorithm, and the EM algorithm. Although the MCMC algorithm is generally able to obtain more accurate estimation results, the algorithm requires a large number of adoptions, is inefficient, and the results are not stable. In the VB algorithm, we need to assume that the parameters and hidden variables in the model are independent of each other, which often leads to the estimation results deviating from the true values and poor learning results but is an effective simplifying computational strategy.

The core idea of the Hardcut EM algorithm is to convert the posterior distribution of samples into a 01 binomial distribution using the maximum posterior probability criterion and then assign the samples to the model components with posterior probability $p = 1$. Due to the distribution characteristics of the data, the majority of the samples have 01 posterior probability distribution, and the error of the hardcut strategy is small in these samples; in the samples at the edges of the model components, the hardcut strategy generates larger errors, but the total error is small due to the small number of samples. On the other hand, the HardcutEM algorithm greatly simplifies the calculation of the Q function in the EM algorithm and improves the speed of the algorithm:

1. Initialization: Use kmeans algorithm to classify sample D into C classes, and initialize hyperparameters (Θ, Ψ);
2. M-step: learning parameters in three steps:
3. Update posterior probability $p\left(z_n^c = 1|x_n\right)$:

$$p\left(z_n^c = 1 \mid x_n\right) = \frac{\pi_c N\left(x_n|\mu_c, S_c^2\right) N\left(y_n|0, \sigma_{1_c}^2 + \sigma_{3_c}^2\right)}{\sum\limits_{c=1}^{C} \pi_c N\left(x_n|\mu_c, S_c^2\right) N\left(y_n|0, \sigma_{1_c}^2 + \sigma_{3_c}^2\right)}. \tag{7}$$

   Update the model parameters $\Psi = \{\mu_c, S_c^2, \pi_c \mid c = 1\}$:

$$\pi_c = \frac{\sum_{n=1}^{N} z_n^c}{N}, \tag{8}$$

$$\mu_c = \frac{\sum_{n=1}^{N} x_n z_n^c}{\sum_{n=1}^{N} z_n^c}, \tag{9}$$

$$S_c^2 = \frac{\sum_{n=1}^{N} \left(\mu_c - x_n\right)^{\mathrm{T}} \left(\mu_c - x_n\right) z_n^c}{\sum_{n=1}^{N} z_n^c}. \tag{10}$$

4. Update the hyperparameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_c \mid c = 1, \cdots, C\}$. The hyperparameters of each GP component are learned independently using a very large likelihood estimation algorithm $\boldsymbol{\theta}_c$.
5. Step E: update the category information of the sample according to the maximum posterior probability principle:

$$z_n^{c^*} = 1. \tag{11}$$

   If $c^* = \arg\max_{1 \ll c \ll C} \{\pi_c N(x_n) N(y_n)\}$.
6. If the change rate of $z_n^{c^*}$ in two iterations is less than the threshold.

3. Clustering Analysis Algorithm

3.1. Algorithm Description

Data types in real databases are complex, and a data object often contains several variables of different types at the same time. It is necessary to process the data before performing calculations. Assume that the data set contains different types of variables and the data matrix is

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \cdots & \cdots & \ddots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nM} \end{bmatrix} \tag{12}$$

To simplify the calculation process, the variables of different types are transformed to a common value space [0.0, 1.0], and the dissimilarity H between objects i and j is defined as

$$d(i, j) = \frac{\sum_{f=1}^{M} \delta_{ij}^{(f)} \cdot d_{ij}^{(f)}}{\sum_{f=1}^{M} \delta_{ij}^{(f)}}, \tag{13}$$

where f represents the variable; x_{if} or x_{jf} represents the metric of the object i or the variable f of the object.

1. When f is a binary or nominal variable, if $x_{if} = x_{jf}$, $d_{ij}^{(f)} = 0$, otherwise $d_{ij}^{(f)} = 1$. If x_{if} or x_{jf} is missing, or $x_{if} = x_{jf}$ and both are asymmetric binary variables, the indicator term $\delta_{ij}^{(f)} = 0$, otherwise $\delta_{ij}^{(f)} = 1$.
2. When f is an ordinal variable, assume that the variable f has V states, corresponding to the sequence V as the rank corresponding to x_{if} , and $r_{ij} \in \left\{1, \cdots, N_f\right\}$, when the weight V can be used instead of x_{if}.
$$Z_{if} = \frac{r_{if} - 1}{N_f - 1}. \tag{14}$$
3. When f is the interval scalar variable, the metric of S_f is standardized and the mean absolute deviation S_f is calculated as
$$S_f = \frac{1}{n} \sum_{i=1}^{n} \left| x_{if} - m_f \right|, \tag{15}$$
where m_f is the average of the f-measure values , i.e.
$$m_f = \frac{1}{n} \sum_{i=1}^{n} x_{if}. \tag{16}$$
Then the normalized metric Z_{if}, is
$$Z_{if} = \frac{x_{if} - m_f}{S_f}. \tag{17}$$

When calculating the phase difference:

$$d_{ij}^{(f)} = \frac{\left| x_{if} - x_{if} \right|}{\max_h x_{hf} - \min_h x_{hf}}. \tag{18}$$

Here h traverses all non-vacant objects of the variable f.

Three kinds of distances are involved in this paper: point-to-point distance; point-to-cluster distance; cluster-to-cluster distance:

1. The distance between points is the most commonly used Euclidean distance, i.e.
$$d(i, j) = \sqrt{\sum_{n=1}^{M} \left(x_{if} - x_{if} \right)^2}. \tag{19}$$
2. The distance between points and clusters is defined as
$$d(i, c) = \min(d(i, j), j \in c). \tag{20}$$
3. The distance between clusters is defined as the average value of the two clusters, with
$$d_{\text{mean}} \left(c_i, c_j \right) = \left| m_i - m_j \right|. \tag{21}$$

Suppose P non-repeating small sample sets $u_1, u_2, \cdots u_P$ are randomly selected in the target database, each small sample set $u_i(1, i, P)$ contains n objects, the number of output classes is K, c_m denotes the cluster of small samples, and $i = r = 1$.

Index project	Index content	Computing method	Remarks
i_1	Total assets	Computing method	Remarks
i_2	assets	after tax / total assets × 100%	Measuring corporate profitability
i_3	Turnover rate of total assets	Product sales revenue / total assets × 100%	Measuring enterprise operating efficiency
i_4	Asset liability ratio	Total liabilities / total assets X100%	Measuring the solvency of enterprises
i_5	Long term debt ratio	Total long-term debt / liabilities × 100%	Measure the long-term solvency of enterprises

Table 1. Indicators of Corporate Bond Credit Ratings

4. Experimental Results and Analysis

4.1. Fuzzy Evaluation Method

Enterprise credit rating is fuzzy, and the influence obtained by using the one-dimensional linear affiliation function, which is called ”single-factor affiliation”, and each indicator is evaluated individually. Secondly, according to the weight of each indicator, the composite operation of the fuzzy matrix is performed on each single factor affiliation to calculate the comprehensive affiliation, and the index value of comprehensive assessment is obtained; Thirdly, the credit status of the enterprise is assessed according to the index value of comprehensive assessment.

Therefore, this paper selects five indicators: total assets, return on assets, turnover rate of total assets, gearing ratio and long-term debt ratio to evaluate the business risks, financial status and debt issuance projects of debt issuing enterprises (as shown in Table 1).

After analyzing the large sample, the optimal, actual, and impermissible values of each indicator are obtained. Assuming that the actual value of the ith indicator of a credit is A_{is}, the standard value (weight) of the indicator is D_i, and A is the composite indicator index of the credit.

When the indicator is positive, the single-factor affiliation d_i of the indicator for this credit is:

$$A_i\left(A_{is}\right) = \begin{cases} 1, & \text{if } A_{is} \sim A_{iy}, \\ \frac{A_{is}-A_{id}}{A_{iy}-A_{id}}, & \text{if } A_{id} < A_{is} < A_{iy}, \\ 0, & \text{if } A_{is}\}A_{id}. \end{cases} \tag{22}$$

When the indicator is an inverse indicator, the single-factor affiliation d_i of the indicator for this credit is:

$$A_i\left(A_{is}\right) = \begin{cases} 0, & \text{if } A_{is} \sim A_{iy}, \\ 1 - \frac{A_{is}-A_{id}}{A_{iy}-A_{id}}, & \text{if } A_{id} < A_{is} < A_{iy}, \\ 1, & \text{if } A_{is}\}A_{id}. \end{cases} \tag{23}$$

When the indicator is an inverse indicator, the single-factor affiliation d_i of the indicator for this credit is:

$$A_i\left(A_{is}\right) = \begin{cases} 0, & \text{if } A_{is} \sim A_{iy}, \\ 1 - \frac{A_{is}-A_{id}}{A_{iy}-A_{id}}, & \text{if } A_{id} < A_{is} < A_{iy}, \\ 1, & \text{if } A_{is}\}A_{id}. \end{cases} \tag{24}$$

The total assets, return on assets, and total asset turnover in this example are positive indicators, while the corporate gearing ratio and long-term debt ratio are inverse indicators, and the composite index is: $A = \frac{\sum_{i=1}^{5} d_i \times D_i}{\sum_{i=1}^{5} D_i}$.

The creditworthiness of a company is assessed based on the value of the indicators of the comprehensive assessment, and the closer the rating result is to 0, the worse the creditworthiness is, and the closer it is to 1, the better the creditworthiness is.

4.2. Results

In the experiment, we first modeled the difference between the 10-year Treasury yield and the 5-year Treasury yield, denoted as ”105”; next, we modeled the difference between the 5-year Treasury

Age	Gender	Marriage	Working conditions	Education	Renewal rate	Loss ratio	Premium amount
0.29	0	1	0.5	0.67	0.8	0.14	0.77
0.14	0	0	0	1	0.83	0.09	0.81
0.14	1	1	1	0.33	1	0.02	0.30
0.43	1	1	1	0.33	1	0	0.18
0.14	0	0	0	0.37	0.67	0.41	0.21
0.71	0	1	1	0.33	0.71	0	0.71

Table 2. Processed Insurance Customer Information Data

yield and the 1-year Treasury yield, denoted as ”5”; and finally, the 10-year Treasury yield is modeled as ”10”. Figure 1 shows the curves of ”105”, CPI, IP, and interbank 7-day pledged repo rate. Since the CPI and IP are updated monthly by the National Bureau of Statistics, the CPI and IP are changed to daily updated values by linear interpolation to maintain consistency.

Based on the form of the data, the paper applies the further improved K-means clustering algorithm to the credit information classification of individual insurance customers. With the help of insurance professionals, some individual customer attributes and business indicators are extracted from the customer information database of a property and casualty insurance company to describe individual customer credit, such as age, gender, education, marital status, employment status, renewal rate, claim rate, and premium amount.

In the experiment, the data from the insurance customer information database for the past two years are selected as the target database, and five small sample sets containing 400 customer information are randomly selected to form the large sample set. The data objects contain various types of variables, which need to be processed before clustering. For example, the age attribute is [20]. For example, the age attribute is divided into 8 intervals such as [20], etc., and the corresponding weight z is calculated with $\{1, 2, \ldots, 8\}$ as the corresponding state value. As the value of the variable, the number of output categories is set to 4. Due to space limitations, Table 2 shows some of the processed data.

The original K-means clustering algorithm and the improved K-means algorithm were used to enhance the efficiency of processing time for large sample sets. The analysis results indicate that the probability values of the differences between categories are less than 0.001, and the clustering effect is good. After clustering the sample data multiple times, the stability of the improved algorithm is 0.795, higher than the original algorithm.

Furthermore, we use the term spread 5-1, the term spread 105, and the 10-year Treasury yield as time series datasets, respectively. Firstly, the time series data are reconstructed using different regression (or recursive) orders and sampling intervals, where the input and output of the reconstructed data are $X_{ti} = \left(X_{ti-d}, X_{ti-2d}, \cdots, X_{ti-pd}\right), Y_{ti} = X_{ti}$, where p is the regression (or recursive) order and d is the sampling interval. Secondly, the Gaussian mixture model with RBF model and SVM regression model are applied to the three datasets. In the experiments, we selected $p = 1, \ldots, 6$, and $d = 1, \ldots, 8$, conducting 48 sets of experiments. Table 2 shows the best experimental results for each algorithm in the 48 sets of experiments and the corresponding p and d. The best experimental results for each algorithm were selected from the 48 sets of experiments on the reconstructed data of the three datasets [21, 22].

From Table 3, we can see that the MGP model obtains the best prediction error RMSE for all three data reconstructions, and we can also observe that the p and d of the reconstructions with the best prediction error differ for different data. This makes it challenging to obtain the optimal p and d in practical applications. Obtaining optimal p and d by model selection algorithms is a promising research direction in the future. In terms of running time, the MGP model still takes the longest time, which is consistent with the results of the first set of experiments.

MGP	Optimal (d,p)	RMSE
5-1	(1,1)	3.64
10-5	(1,6)	3.46
10	(1,1)	3.68
SVM	Optimal (d,p)	RMSE
5-1	(1,1)	5.22
10-5	(1,1)	3.91
10	(4,5)	2.85
RBF	Optimal (d,p)	RMSE
5-1	(1,1)	4.49
10-5	(2,1)	3.59
10	(1,1)	1.94

Table 3. Results of Regression Analysis of Each Algorithm on Three Sets of Recombination Data

5. Conclusions

The exploration and study of corporate credit term structures have garnered significant attention due to their substantial value in corporate credit analysis and market investment. This topic has become a crucial area in financial engineering, attracting scholars and investors alike. This paper initiates an analysis of domestic and international approaches to interest rate term structures. It observes that existing studies are limited to exploring the characteristics of interest rate term structures based on known market behavior. By delving into a substantial amount of historical data, this paper identifies three key factors influencing the term structure of government bond interest rates: the inflation index CPI, the growth rate of industrial value added IP, and a crucial measure of market funding-the interbank 7-day pledged repo rate. Breaking away from the traditional academic thinking framework, this paper employs a Gaussian process mixture (MGP) model to predict future behavior effectively. This approach considers market participants' perspectives while respecting historical changes in the market. Experimental results demonstrate that the MGP model achieves more accurate prediction results compared to other machine learning algorithms. It also exhibits a significant advantage over traditional linear regression algorithms in capturing market dynamics.

References

1. Petropoulos, A., Chatzis, S. P., and Xanthopoulos, S., 2016. A novel corporate credit rating system based on Student'st hidden Markov models. *Expert Systems with Applications, 53*, pp. 87-105.

2. AlMahmoud, R. H., Hammo, B., and Faris, H., 2020. A modified bond energy algorithm with fuzzy merging and its application to arabic text document clustering. *Expert Systems with Applications, 159*, p. 113598.

3. Bhatnagar, V., Majhi, R., and Jena, P. R., 2018. Comparative performance evaluation of clustering algorithms for grouping manufacturing firms. *Arabian Journal for Science and Engineering, 43*(8), pp. 4071-4083.

4. Mohammadi, M. G., Mahmoud, D., and Elbestawi, M., 2021. On the application of machine learning for defect detection in L-PBF additive manufacturing. *Optics & Laser Technology, 143*, p. 107338.

5. Pai, G. V., and Michel, T., 2009. Evolutionary optimization of constrained $ k $-means clustered assets for diversification in small portfolios. *IEEE Transactions on Evolutionary Computation, 13*(5), pp. 1030-1053.

6. Tayal, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T., and Tyagi, N., 2015. Crime detection and criminal identification in India using data mining techniques. *AI & society, 30*(1), pp. 117-127.

7. Golbayani, P., Florescu, I., and Chatterjee, R., 2020. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance, 54*, p. 101251.

8. Vignesh, B., Oliver, W. C., Kumar, G. S., and Phani, P. S., 2019. Critical assessment of high speed nanoindentation mapping technique and data deconvolution on thermal barrier coatings. *Materials & Design, 181*, p. 108084.

9. Tarassenko, I., and Roberts, S., 1994. Supervised and unsupervised learning in radial basis function classifiers. *IEE Proceedings-Vision, Image and Signal Processing, 141*(4), pp. 210-216.

10. Lu, Y., and Christou, A., 2019. Prognostics of IGBT modules based on the approach of particle filtering. *Microelectronics Reliability, 92*, pp. 96-105.

11. Zheng, Z., Chen, B., Xu, Y., Fritz, N., Gurumukhi, Y., Cook, J., Ates, M.N., Miljkovic, N., Braun, P.V. and Wang, P., 2021. A Gaussian process-based crack pattern modeling approach for battery anode materials design. *Journal of Electrochemical Energy Conversion and Storage, 18*(1), p.011011.

12. Lin, R. H., Xi, X. N., Wang, P. N., Wu, B. D., and Tian, S. M., 2019. Review on hydrogen fuel cell condition monitoring and prediction methods. *International Journal of Hydrogen Energy, 44*(11), pp. 5488-5498.

13. Song, Y., Wang, Y., Ye, X., Wang, D., Yin, Y., and Wang, Y., 2020. Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. *Information Sciences, 525*, pp. 182-204.

14. Protopapadakis, E., Niklis, D., Doumpos, M., Doulamis, A., and Zopounidis, C., 2019. Sample selection algorithms for credit risk modelling through data mining techniques. *International Journal of Data Mining, Modelling and Management, 11*(2), pp. 103-128.

15. Zhang, S., Xiong, W., Ni, W., and Li, X., 2015. Value of big data to finance: observations on an internet credit Service Company in China. *Financial Innovation, 1*(1), pp. 1-18.

16. Ruan, T., Lei, L., Zhou, Y., Zhai, J., Zhang, L., He, P., and Gao, J., 2019. Representation learning for clinical time series prediction tasks in electronic health records. *BMC medical informatics and decision making, 19*(8), pp. 1-14.

17. Brusco, M. J., Steinley, D., Cradit, J. D., and Singh, R., 2012. Emergent clustering methods for empirical OM research. *Journal of Operations Management, 30*(6), pp. 454-466.

18. Zhou, T., Song, Z., and Sundmacher, K., 2019. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. *Engineering, 5*(6), pp. 1017-1026.

19. Zhou, J., Sun, J., Zhang, W. and Lin, Z., 2023. Multi-view underwater image enhancement method via embedded fusion mechanism. *Engineering Applications of Artificial Intelligence, 121*, p.105946.

20. Zhou, J., Sun, J., Zhang, W. and Lin, Z., 2023. Multi-view underwater image enhancement method via embedded fusion mechanism. *Engineering Applications of Artificial Intelligence, 121*, p.105946.

21. Vladimir, M., Matwijczuk, A.P., Niemczynowicz, A., Kycia, R.A., Karcz, D., Gladyszewska, B., Slusarczyk, L. and Burg, P., 2021. Chemometric approach to characterization of the selected grape seed oils based on their fatty acids composition and FTIR spectroscopy. *Scientific Reports, 11*(1), p.19256.

22. Nystrup, P., Kolm, P. N., and Lindström, E., 2020. Greedy online classification of persistent market states using realized intraday volatility features. *The Journal of Financial Data Science, 2*(3), pp. 25-39.