Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction Poverty Levels of College Students Using a Machine Learning Model

Wang Sheng (✉ ws@chzu.edu.cn)
Chuzhou University
Shi Yumei
Chuzhou University

Research Article

Keywords: poverty levels, Targeted poverty alleviation, Feature selection, Feature extraction

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-919541/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Prediction poverty levels of college students Using a machine learning model

Wang Sheng[1] and Shi Yumei[2*]

[1]Center of information development and management, Chuzhou University, HuifengRoad, Chuzhou, 239000, Anhui, China.
[2*]School of mathematics and finance, Chuzhou University, HuifengRoad, Chuzhou, 239000, Anhui, China.

*Corresponding author(s). E-mail(s): sym@chzu.edu.cn;
Contributing authors: ws@chzu.edu.cn;

Abstract

Nowadays, poverty-stricken college students have become a special group among the college students and occupied higher proportion in it. How to accurately identify poverty levels of college students and provide funding is a new problem for universities. In this manuscript, a novel model that combined Random Forest with Principle Components Analysis (RF-PCA) is proposed prediction poverty levels of college students. To build this model, data was firstly collected to establish datasets including 4 classed of poverty levels and 21 features of poverty-stricken college students. Then, feature dimension reduction includes two steps: the first step we selected the top 16 features with the ranking of feature, according to the Gini importance and Shapley Additive explanations (SHAP) values of features based on Random Forest (RF); the second step of feature extraction through Principle Components Analysis (PCA) extracted 11 dimensions. Finally, confusion metrics and receiver operating characteristic (ROC) curves were used to evaluate the performance of the proposed model, the accuracy of the model achieved 78.61%. Furthermore, compared with seven different classification algorithms, the model has a higher prediction accuracy, the result has great potential to identify the poverty levels of college students.

Keywords: poverty levels, Targeted poverty alleviation, Feature selection, Feature extraction

1 Introduction

In 2013, China put forward the concept of targeted poverty alleviation [1]. To fundamentally address the problem of poverty-stricken college students, the country, society and schools actively explore and initially establish a funding system for poor students. In 2019, government, universities and society formulated various financial aid policies for college students, providing a total of 48.1759 million students national wide. The targeted funding model of universities is a concrete practice that embodies the idea of targeted poverty alleviation [2]. Therefore, accurate identification the poverty level is of great significance to improving quality of poverty alleviation .

Despite targeted poverty alleviation great importance, poverty prediction or classification is labor-intensive and time-consuming in developing countries. In principle, the work of identifying students from family financial straits should be carried out once every academic year, Anhui province. The procedures for the identification mainly include four steps: notice in advance, individual application, college appraisal and result public . However, there are three additional factors that could affect the identification results in practice:

(1) To protect personal privacy, providing false information in the application form [3];

(2) On other sessions, the reviewers (peers or teachers) may be influenced by subjective factors.

(3) Different affiliated college has different implementation standards.

How to apply new technologies thinking to the financial aid for poor students in colleges and universities requires us not only to change the thinking of financial aid, but also to use advanced technology to improve and innovate the traditional methods.

A supervised machine learning algorithm is widely accepted for regression or classification in recent decades. Regression, i.e., predicting a new value based on existing values, and classification, i.e., predicting the outcomes by the existing culverts. The learning algorithm can be a single model or an ensemble model and could be linear or nonlinear. Machine learning has driven advances in many different fields and has shown great potential in predicting poverty.

In this paper, the RF-PCA model was proposed. The overall implementation process of the model is shown in Figure 1, and the whole process include three stages: dataset collection, model construction and model evaluation. Firstly, collected the original poverty-stricken college student dataset contains two sections, the labels are from the result of college identification, and the features are from Student Information Management System (SIMS). Then, the feature selection, which was used for Gini-importance and SHAP values based on RF, and the feature extraction, which was employed for PCA. Finally, the RF-PCA model was established by using the extracted feature data. The performance of the RF-PCA model was evaluated via confusion matrix and ROC curve, and compared with seven other popular classification algorithms.

The main contributions of this work are listed as follows.

(1) Put forward the main factors affecting poverty

Redundant and less important attributes may weaken the results of the model, removing redundant features can avoid the overfitting, speed up model training.

(2) Establish the recognition of the model

The proposed model has achieved an identification accuracy of 78.61%, contribute to simplify the traditional methods and eliminate the subjective factors.

(3) Protect the poverty-stricken students' privacy

Using the model to identify the poverty levels, students do not need to fill in the application form and worried about privacy being revealed. The model directly analyzes the original data, which can well protect the privacy of students and eliminate their concerns.

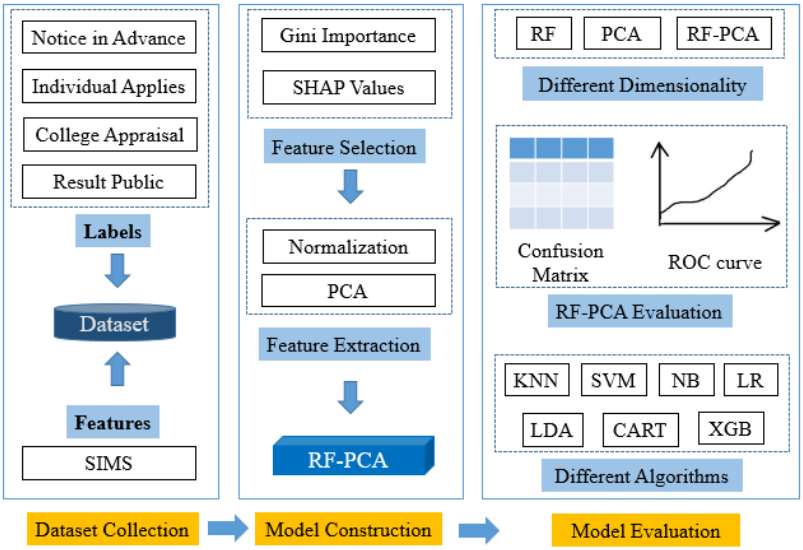


Fig. 1 Overview workflow of RF-PCA

2 Related work

The financial aid policies of university students have become an important part of the international education research subject. Shi et al. [3] establish a subsidy system for colleges by improving the information value, docking precise funding data-link, develop funding information files card, sharing funding information, and dynamic information supervision. Bai et al.[4] analyzed and evaluated the effect of financial aid based on big data. Shi et al. [3], mentioned the enlightenment of big data thinking on the identification of poverty-stricken students. These theoretical studies provide a direction for technical studies.

The availability of data has improved substantially in recent years, the researchers have started to utilize big data and machine learning to identify poverty levels.Big data technology is one of the most popular of prediction poverty levels. For each student generate huge quantities of data during the school, the emergence of big data provides a new way to solve these problems [5]. Cao et al. [6] and Xu et al. [7] analyzed students' consumption data in school canteens based on big data. Yang et al. [8] based on the [6] and [7] literature, increased data types included library data, bedroom door access data and performance data. Hence, the characteristic of big data is to collect a lot of information about students during the academic year, and needs large-scale data center for data storage, but some college without a data center.

Some researchers pointed out students' habitual behaviors and life traces which truly and objectively reflect students' daily life, these can be analyzed and evaluated the financial situation [9]. Tao et al. [10] employed students' daily lives data to build the simulation modeling for the poverty-stricken college students. Wu et al. [11] and Chao et al. [9] proposed build a smart campus system based on campus big data to identify poverty-stricken students. However, collecting information may interfere with students' normal life, making hard to use in practice.

Machine learning techniques have been widely used in in different fields, targeted poverty reduction can be further assisted by technology.Mohamud et al. [12] designed an approach to identify poverty levels that extract a subset of feature, examine this subset affect and employ ensemble models, which based partly on machine learning techniques.

The proxy means test (PMT) is a commonly used targeting poverty tool by employing the distinguished feature of the household when proof of income is not provided [13]. Mcbridea et al. [14] presented evidence that machine learning algorithm was used for PMT development can adequately enhance the out-of- sample capability of targeting tools. Sani et al. [15] used machine learning to choose the best identify bottom 40 percent household population. These studies suggested an excellent performance by decision tree models.

An intelligent evaluation model for poor college students based on C4.5 decision tree algorithm was established by the historical data of student information and types of poverty [16]. In the work, machine learning method was utilized to confirm the respective strength of each explaining attributes of escaping poverty and falling into poverty [17]. Irawan et al. [18], compared the accuracy of k-nearest neighbor (KNN) and Learning vector quantization (LVQ) in poverty levels classification, the result manifested that the KNN showed better performance than LVQ.

Back-propagation (BP) neural network algorithm has also been applied in the field of poverty. Shao et al. [19] constructed the BP neural network, and created a nonlinear mapping between the identification of poverty-stricken students and the economic status of college students. The poor households and non-poor households were predicted using BP neural networks [20] . Some researchers combine satellite imagery and machine learning, especially deep learning to analyze poverty problems [21–23].

Although the above methods have excellent performance on poverty problem, which difficult to widely used in college. Unlike these studies, data sources for our model only needs to be filled in by students alone once in SIMS, and reduce infrastructure requirements.

3 Methods

3.1 Data normalization

Data normalization aims to the data fall into the same interval, in order to better adapt machine learning algorithm. Different normalization methods maybe impact the prediction results. The Min-Max and Z-score are common normalization methods [24, 25]. Min-Max standardization could normalize all feature values to interval $[0, 1]$ or interval $[-1, 1]$. These standardized formulas are as follows:

$[0, 1]$ normalization:

$$X_{[0,1]} = \frac{X - X_{Min}}{X_{Max} - X_{\mathrm{Min}}} \tag{1}$$

$[-1, 1]$ normalization:

$$X_{[0,1]} = \frac{X - X_{Min}}{X_{Max} - X_{\mathrm{Min}}} \tag{2}$$

Z-score normalization:

$$X_Z = \frac{X - \mu}{\sigma} \tag{3}$$

Here, X denotes the initial sample data; X_{Max}, X_{Min}, μ and σ denote the maximum, minimum, average and standard deviation values of the feature, respectively.

3.2 Feature importance

3.2.1 Gini-importance

The feature importance score is based on the Gini index, built into the scikit-learn implementation of RF [26]. A higher feature importance score indicates a more important feature, which has a larger effect on the model. The Gini index is calculated as follows:

$$\mathrm{Gini} = \sum_{i=1}^{n} -f_i\left(1 - f_i\right) \tag{4}$$

Here, f_i denotes the frequency of a label at a node, and n denotes the number of labels. Multiple decision trees constitute an RF. The importance

n_j of a node j in each decision tree is represented by Gini impurity:

$$n_j = w_j C_j - \sum_{1}^{m} w_{m(j)} C_{m(j)} \tag{5}$$

Here, w_j denotes the weighted number of samples reaching node j, C_j denotes the impurity value of node j, and m denotes the number of child nodes of the tree. The importance of feature i on decision tree is computed as:

$$f_i = \frac{\sum_1^s n_j}{\sum_{k \in allnodes} n_k} \tag{6}$$

Here, s denotes the times of node j split on feature i. The standardized feature importance in a decision tree is computed as:

$$f_i' = \frac{f_i}{\sum_{j \in \text{ all features in a tree } f_j}} \tag{7}$$

The ultimate importance score of a feature in RF is computed as

$$F_i = \frac{\sum_{j \in \text{ all trees }} f_i'}{N} \tag{8}$$

Here, $\overset{'}{f}$ denotes the standardized feature importance values of a decision tree, N denotes the total number of decision trees [27–31].

3.2.2 SHAP values

SHAP is a unified framework for interpreting predictions of machine learning model [32, 33]. It uses the Shapley values from game theory to estimate contribution to the prediction [34, 35]. For the RF model, SHAP package can be utilized to calculate “SHAP values” for each feature.

3.3 Principle Components Analysis

PCA is a type feature extraction technique that projects the original feature matrix into a new space with lower dimension while retaining most of the available information [36]. In the implementation of PCA algorithm, the calculation of principal component is realized by diagonalization of the covariance matrix of the data. The relevance of the principal components can be ranked in terms of the eigenvalues, and the variance of each principal component reflects its contribution to the data. Supposed that the original sample data matrix $\boldsymbol{X}^{m*n}$ is in a form of m rows and n columns.

The PCA algorithm achieves this as follows:

Step 1: Read data matrix $\boldsymbol{X}$;

Step 2: Each column represents a feature, and the mean value is obtained. The mean value is subtracted from the initial data to the new centralized data;

Step 3: The covariance matrix is calculated:$D(\boldsymbol{X}) = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$;

Step 4: Eigenvalue decomposition approach is used to compute eigenvalue λ and eigenvector q of the covariance matrix;

Step 5: The eigenvalues are ranked from high to low, and extract the first k of them, the eigenvector $\boldsymbol{W}$ is defined by the k eigenvectors;

Step 6: Multiply the data set $m * n$ by the eigenvector of n dimensional eigenvector, and obtain the data matrix $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}$.

In many studies, cumulative variance explained contribution rate more than 85% were used as a basis for extracting k principal components, which can represent the features of the initial sample data [37–39].

3.4 Statistical analysis of multi-class predictions

For binary classification, a confusion matrix consist of elements: true positive (TP), false positive (FP), false negative (FN) and true negative (TN) [40]. Evaluation metrics performance of classification algorithms, including accuracy (ACC), precision (PR), recall (RE) and F1-score (F1), which can be computed by confusion matrix. These calculation formulas are as follows:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$PR = \frac{TP}{TP + FP} \tag{10}$$

$$RE = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \times PR \times RE}{PR + RE} \tag{12}$$

Both micro-average and macro-average is adopted to evaluate the performance of multi-classification problem, we need to conceptualize the problem as a binary classification problem by using One vs Rest [41]. Micro-average metric constitutes a mean biased by class frequency. It can compute the micro-average precision (*micro*-P), micro-average recall (*micro*-R) and micro-average F1-score (*micro*–F1), respectively. The formula as follows:

$$\text{micro } - P = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{13}$$

$$\text{micro } - R = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{14}$$

$$\text{micro } - F1 = \frac{2 \times \text{ micro } - R \times \text{ micro } - P}{\text{micro } - R + \text{ micro } - P} \tag{15}$$

If the confusion matrix were a square matrix, the *micro*–P, *micro*–R and *micro*–F1 would be equal, we will use *micro*-Avg unified representation.

All classed are considered equally important in the macro-average calculation. Macro averaging for n classes compute the metrics individually for each class and average results together [42]. It can compute the macro-average precision (*macro*-P), *macro*-average recall (*macro*-R) and macro-average F1-score (*macro*–F1) respectively. The formula as follows:

$$\text{macro} - P = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i} \tag{16}$$

$$\text{macro} - R = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} \tag{17}$$

$$\text{macro} \quad - F1 = \frac{2 \times \ \text{macro} \quad - R \times \ \text{macro} \quad - P}{\text{macro} \quad - R + \ \text{macro} \quad - P} \tag{18}$$

Here, n denotes the number of classes in the specific classification problem; TP_i denote the number of samples that are correctly classify ith into *ith* class, $TP_i = T_i P_i$; FP_i denote the number of samples that are wrongly classify *jth* class into *ith* class; FN_i denote the number of samples that are wrongly classify *jth* class into other classes;

The higher the Equations (9)-(18) value, the better the model performance.

4 Experiments

4.1 Datasets

4.1.1 Data description

To obtain the data and build the prediction model for poor students, we collected features of poor college students at the Chuzhou University from Student Information Management System (SIMS). Chuzhou University, founded in 1950 and is located in Anhui Province, China. It is a full-time undergraduate university with an enrollment of more than 19000 students, about a quarter of all students were funded. The poverty-stricken students been divided into four different types of poverty, and each poverty-stricken college student has 21 attributes. The 21 features were filled in by students in the SIMS independently, include 6 continuous attributes and 15 categorical attributes. Table 1 demonstrates the poverty levels and features of poverty-stricken students.

4.1.2 Data pre-processing

For best use the poverty-stricken college student dataset, we need to preprocess the original data. 5000 samples from Chuzhou University were collected. The dataset had missing values because some students who decide not to answer all questions considering their privacy; Partial eigenvector values were distinctly wrong, due to misunderstanding column names. The missing values and error values were removed by filtering, which the number of samples decreased to 4604. One-hot encoding was utilized to deal with categorical features,by which a new binary feature was generated for each level of each categorical feature

Table 1 The poverty levels and features of poverty-stricken students

Categories	Variables
Poverty levels (n=4)	Slightly weak difficult (SWD), Weak difficult (WD), Slightly strong difficult (SSD), Strong difficult (SD)
Features(n=21)	Amount of debt (AOD), Per capita annual household Income (PCAHI), The number of family unemployment (TNOFU), Labor force (LF), The number of dependents (TNOD), The number of family members (TNOFM),Rural beneficiaries (RB), Subsistence allowance (SA), 'Orphan or not' (OON), Single parent child (SPC), Children of disabled persons (CODP), 'Disabled or not?'(DON), Serious illness of family members (SIOF), Parents lose the working capability (PLTWC), Poverty registration (PR), Low-income family(LIF), Children of military (COM), Suffered from natural disasters (SFND), Suffered from accidents(SFA), Rural subsistence(RS), Rural special poverty support (RSPS)

[43]. The categorical features were all literal, we created a new binary feature. And then, in order to the values fall into the same numeric interval, all the variables were normalized.

4.2 Feature selection

4.2.1 Parameter Tuning

To obtain the parameters of the best model, the parameters of RF algorithm need to be adjusted before feature selection. For the RF model, the following mainly hyper-parameters were tuned:n_estimators, max_depth, min_samples_split, and min_samples_leaf. Experiments utilizing a grid search to fine tune the trade-off between the metrics on the testing set and the perplexity of the model, engaging significant predictions and avoiding overfitting. The model was trained and evaluated on the dataset using 10-fold cross-validation (CV). Finally, we used the following parameter settings: n_estimators = 61, max_depth = 16, min_samples_split = 6, min_samples_leaf = 11, and default values of the other remaining parameters were utilized. After all parameter tuning, the RF model achieved an accuracy of 75.56%.

4.2.2 Estimation of feature importance

There are 21 features in the initial poor student dataset, each of which contains relevant information and play a different role in build model process. After training, the ranking of the feature importance score based on Gini importance for RF model are shown in Figure 2. The 21 features were sorted on the basis of the order of the importance scores from high to low. It can see the AOD achieved the highest score of 0.2539, which indicated AOD feature was likely to the most important; In contrast, the COM achieved the lowest score indicated the least important. Besides, other features achieved a different score, but we need to focus on low-scoring features, for example, the importance score of DON, OON, CODP, RB were less than 0.01, which maybe belong to redundancy features.

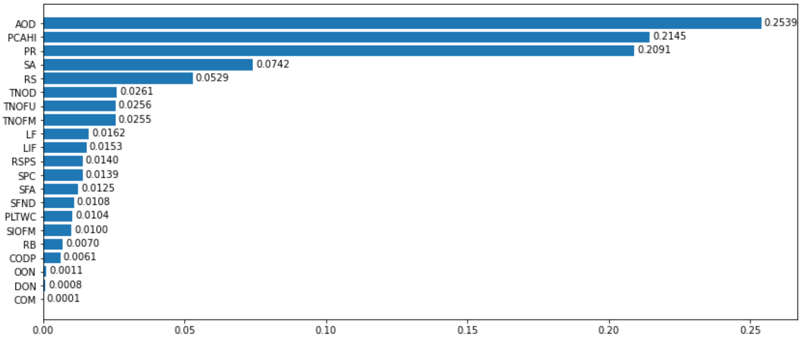


Fig. 2 Rank the 21 features by importance scores

4.2.3 Model interpretation by SHAP

Next, SHAP were used to further interpret the trained RF model. Figure 3 is the bar graph based on the mean absolute SHAP value for each feature. Here, the higher the SHAP values, the more important the feature. It can see that PR had the biggest impact on model output, while COM had less impact. Meanwhile, the influence of DON and OON on the model output was close to zero. Compared Figure 2 with Figure 3, we can see that the five features were the same at the bottom, they were respectively COM, DON, OON, CODP and RB, but changed from the sixth from the bottom. To eliminate redundant features, we step by step excluded the bottom 5 features in Figure 3 from the dataset and repeated training the RF model. Table 2 shows prediction accuracy for different the number of the remaining features.

Table 2 Different numbers of features for accuracy

Remove features	The number of remaining features	Accuracy
COM	20	76.14%
COM, DON	19	77.47%
COM, DON, OON	18	77.88%
COM, DON, OON, CODP	17	77.86%
COM, DON, OON, CODP, RB	16	77.87%

Obviously, after removing the COM, the model accuracy ascent from 75.56% to 76.14%, and when DON, OON, CODP, RB were removed, overall accuracy have been further improved. Compared with the results from original dataset based on RF model, after removing the five features had no significant impact on the accuracy of the RF model, this indicated that they were

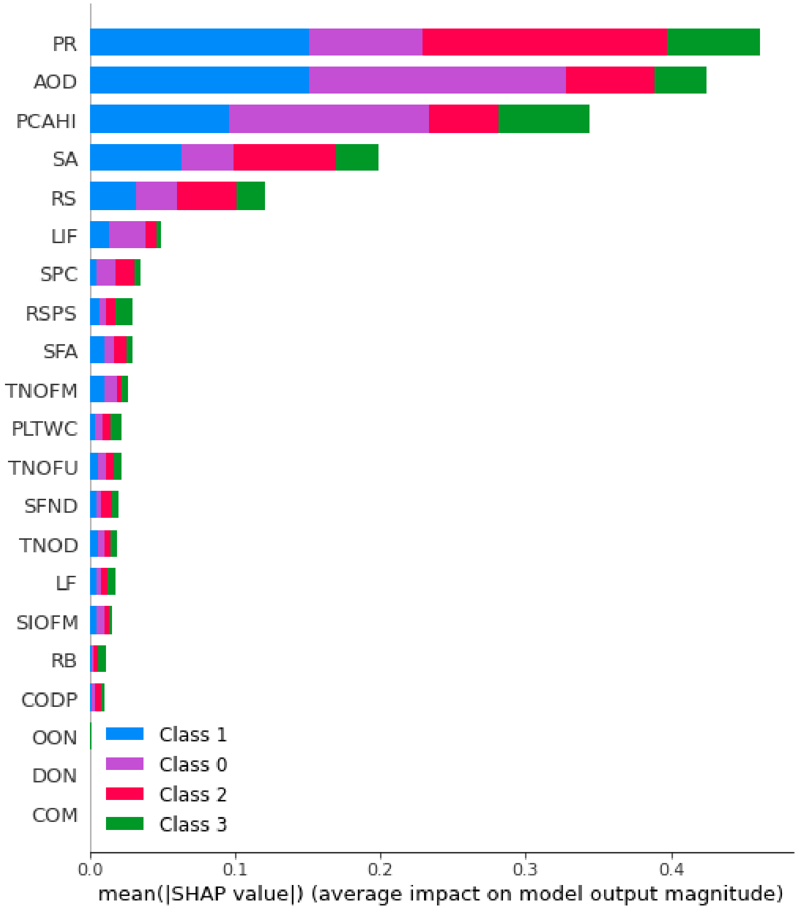


Fig. 3 The SHAP values of Feature information score

redundancy features. Finally, remaining 16 features were selected as next step of research.

4.3 Feature extraction

4.3.1 Different normalization based on PCA

PCA was employed to extract the 16 features after feature selection. To remove the impacts of excessive dimensional difference between features, before using PCA for extracting features, 16 features were further normalized Z-score and [0, 1] standardization methods. In this study, to remain more characteristic information and ensure the reliable of the model, we set goals that the cumulative contribution rate was more than 85%.

Under two different data standardized methods, the input the data of feature extraction for the prediction model. The corresponding relationship between the cumulative explained variance ratio and the accuracy rate in different dimensions was compared respectively as shown in Figure 4. It can be seen that different normalization method leads to different rates of convergence. For the cumulative explained variance ratio, the [0, 1] converges faster than the Z-score, [0, 1] begin to converge when the number of components was greater than 10, but the z-score begin to converge until the last two. For prediction accuracy, the Z-score converges faster the [0, 1], while there was an obvious difference in the process.

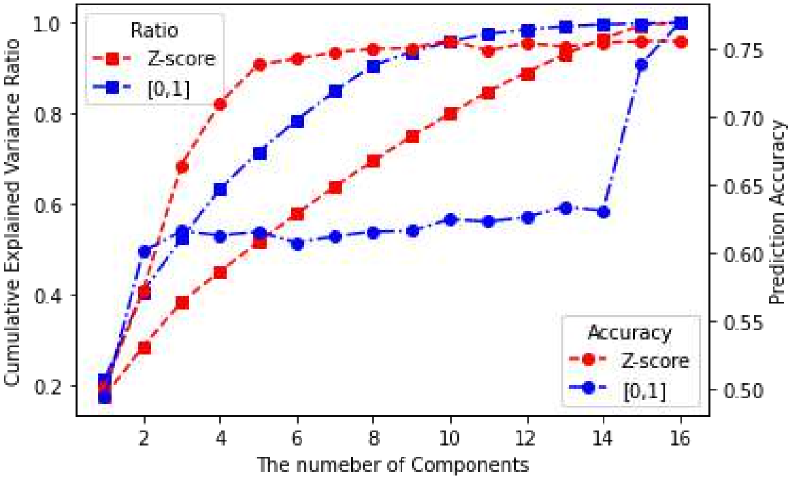


Fig. 4 The different number of components cumulative explained variance ratio and prediction accuracy.

By comparing and analyzing the cumulative contribution rate and predictive results under different normalization methods. The conclusion obtained that the cumulative contribution rate of the first 11 principal components was 85% by Z-score standardization, and the higher accuracy of the model, the result of feature extraction was the selection of the first 11 components. Finally, Z-score was selected as the normalization approach, and after the PCA dimension reduction, the feature dimension was changed from 21 to 11.

4.3.2 Performing PCA of the original data

PCA was used to individually examine the relationship between the cumulative contribution rate of the original data and single interpretation variance in Figure 5. It can be seen that the first 15 principal components reached 85.5% for the cumulative contribution rate, completed more than the required 85%. Afterwards, the 15 dimensions through PCA extracted as input trained model also based on RF model, in order to this process facilitate the marking, simply as PCA-RF. To further verify the plausibility of feature dimensionality reduction order, four different dimensionality of data for prediction accuracy were

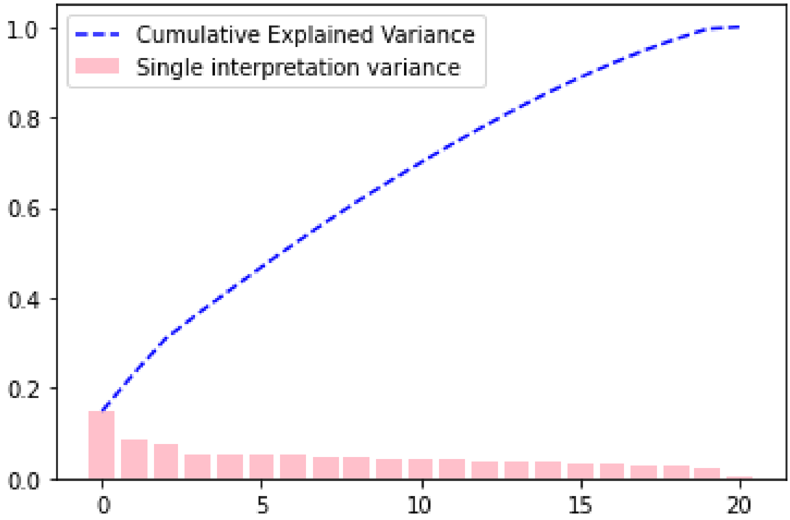


Fig. 5 Eigenvalues and cumulative contribution ratio of principal components

compared based on RF model, including the 21 dimensional initial data, the16 dimensional data after direct feature selection, the 15 dimensional data after direct feature extraction, and the 11 dimensional data after feature selection and feature extraction. The results were shown in Table 3, the RF-PCA model used only 52.4% of the original feature number, and achieved an accuracy of 78.61%, which were 2.31%, 1.06%, and 3.05% higher than those gained by other methods of dimensionality reduction. This suggests that this approach to dimensionality reduction was reliable.

Table 3 Prediction accuracy of different dimensionality

Methods	Features	Accuracy
RF	21	75.56%
RF	16	77.87%
PCARF	15	76.5%
RFPCA	11	78.61%

5 Model Evaluation

5.1 Confusion Matrix

To evaluate the RF-PCA model performance, we used a confusion matrix of the classification results to compute the performance indicators and analyze the misclassifications for each class. Figure 6 shows a confusion matrix of $4 * 4$ for the predicted results of poverty levels, which was a square matrix. Each column and row of the matrix represents the prediction and actual labels, respectively [44]. Furthermore, in each cell the numbers denote counts of the number of levels, empty cells means 0. In each cell color was used to depict the percentage of actual labels individuals predicted within each label category (0% indicated by pure white, and then color regularly changing to deep black indicating 100%).

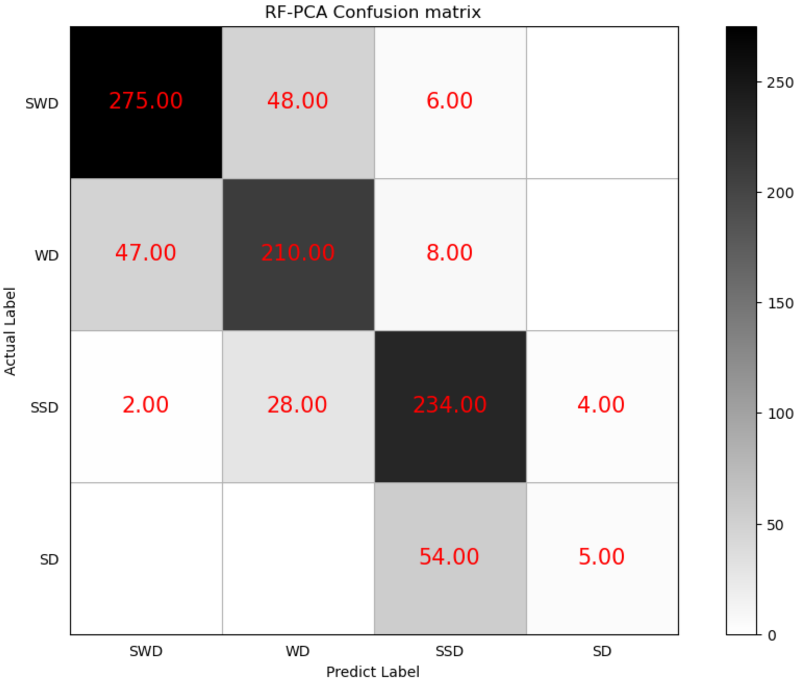


Fig. 6 Confusion matrix for the predicted result of four poverty levels

In the confusion matrix in Figure 6, diagonal values were the number of correct prediction results, while off-diagonal show the number of misclassification. It can calculated the PR, RE and F1 for each label, as listed in Table 4, by the digit of each cell. The results show the RF-PCA model yielded better classification performance for the SWD, WD and SSD, a small amount of confusion happen between adjacent categories. However, the SD was almost completely unpredictable, as 91.53% of the SD were misclassified into SSD, the evaluation metrics of SD (ACC=0.085, PR=0.5556, RE=0.0847, F1 =0.1471) were all the lowest.

5.2 ROC Curve

The ROC curve graph shows the performance of the RF-PCA model. The area under the receiver operating characteristic curve (AUC-ROC) gives the

Table 4 Report for the classification of four poverty levels

Levels	ACC	PR	RE	F1
SWD	0.8359	0.8488	0.8359	0.8423
WD	0.7925	0.7343	0.7925	0.7623
SSD	0.8731	0.7748	0.8731	0.8211
SD	0.0847	0.5556	0.0847	0.1471

performance index of the classifier. The higher the AUC values, the better the prediction of the model.

Figure 7 shows, that the average per-level ROC curves for each level was calculated for the test set by confusion matrix, in order to observe the difference between them. The average per-level ROC curves manifest that level SWD and level SSD were better identified by the RF-PCA model than level WD and level SD. The same results were reflected also by computing the AUC values for the average ROC curves of the four levels (0.96 for level SWD, 0.92 for level WD, 0.95 for level SSD and 0.89 for level SD).

Figure 7 also shows the greater variability of the ROC curves relative to level SD, compared to those obtained with level SWD, level WD, and level SSD, the average ROC curves also confirm the calculated results of Table 4 for each level. Figure 8 shows that the difference between the micro-averaged and macro-averaged ROC curves for the RF-PCA model. The AUC-ROC indicates

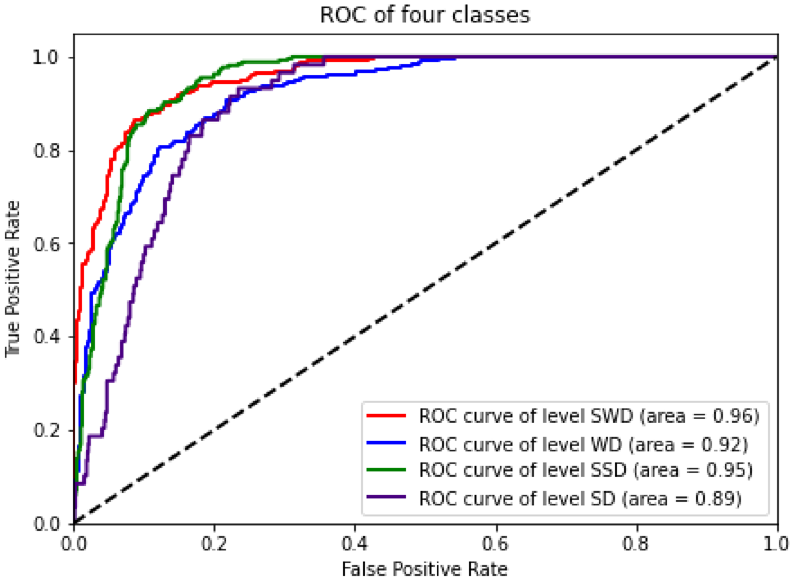


Fig. 7 The average per-level ROC curves calculated

the performance of the overall model, an AUC-ROC were 0.95, 0.93 micro-averaged and macro-averaged forms, both of which were quite reliable.

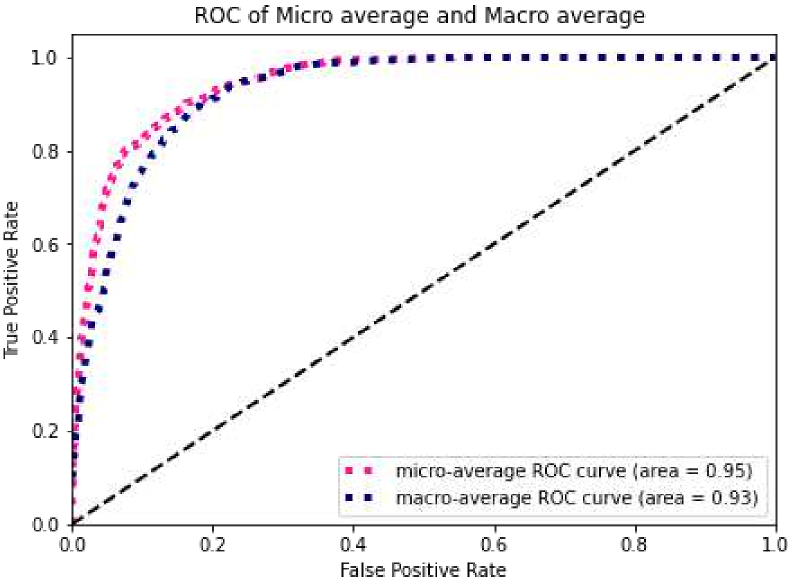


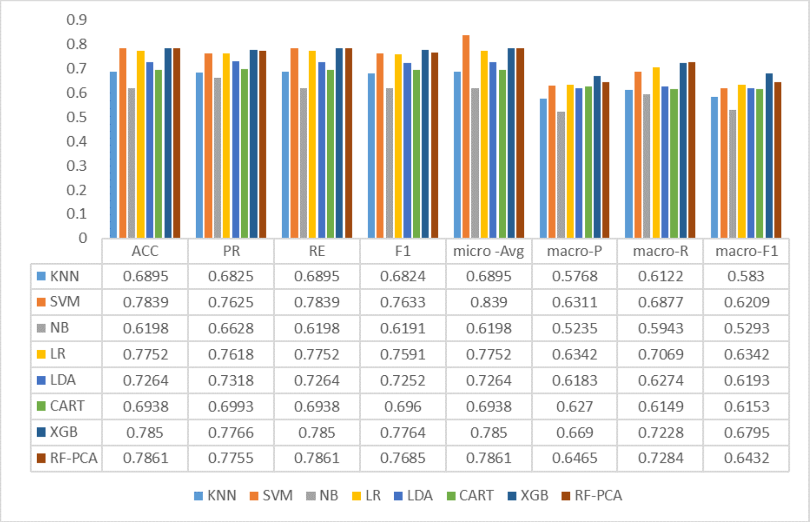
Fig. 8 The ROC curves differences between the micro-averaged and macro-averaged AUCROC

5.3 Different Classification Algorithms

To verify the superiority of the RF-PCA dimensionality reduction, seven different classification algorithms were applied individually, including KNN, Support Vector Machine (SVM), GaussianNB (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Classification And Regression Tree (CART), and extreme Gradient Boosting (XGB) [45]. Each algorithm used the grid search method to optimize hyper-parameters, and divided the data set into 80% training (N=3682) and 20% (N=921) help testing.

The classification capabilities of different algorithms were compared as shown in Figure 9. It can be found that the output results of the seven algorithms were different. The RF-PCA model achieved an ACC of 0.7861, PR of 0.7755, RE of 0.7861, which was the highest among all algorithms. The performance of the XGB algorithm was slightly worse than that of RF-PCA (i.e., ACC = 0.7850, PR=0.7766, RE=0.7764, F1=0.7685,).

In addition, the macro-averaging (*macro*-P, *macro*-R, *macro*-F1) and micro-averaging (*micro*-P, *micro*-R, *micro*-F1) were also utilized to estimate the overall capability of the seven classification algorithms. The micro-average (*micro*-P=*micro*-R=*micro*-F1=0.7861) and macro average (*macro*-P=0.6465, *macro*-R=0.7284, *macro*-F1=0.6432) of the RF-PCA model was slightly higher, respectively. In details, each algorithm has the same values of ACC, RE, and micro-Avg, due to the confusion matrix was a square matrix of $4 * 4$. Finally, the results show that the RF-PCA model had better performance.



	ACC	PR	RE	F1	micro -Avg	macro-P	macro-R	macro-F1
KNN	0.6895	0.6825	0.6895	0.6824	0.6895	0.5768	0.6122	0.583
SVM	0.7839	0.7625	0.7839	0.7633	0.839	0.6311	0.6877	0.6209
NB	0.6198	0.6628	0.6198	0.6191	0.6198	0.5235	0.5943	0.5293
LR	0.7752	0.7618	0.7752	0.7591	0.7752	0.6342	0.7069	0.6342
LDA	0.7264	0.7318	0.7264	0.7252	0.7264	0.6183	0.6274	0.6193
CART	0.6938	0.6993	0.6938	0.696	0.6938	0.627	0.6149	0.6153
XGB	0.785	0.7766	0.785	0.7764	0.785	0.669	0.7228	0.6795
RF-PCA	0.7861	0.7755	0.7861	0.7685	0.7861	0.6465	0.7284	0.6432

Fig. 9 Different algorithms classification metrics

6 Conclusion

In this study, a recognition poverty level method based on feature selection and extraction was proposed, namely RF-PCA. Firstly, the two steps to perform feature selection were ranking of feature importance of Gini importance, and model interpretation of SHAP values. Secondly, PCA was used for feature extraction after feature selection, and the prediction accuracy of different input dimensions using RF, PCA-RF, RF-PCA alone were compared. Finally, the ROC curve and the performance indicators were used to evaluate the performance of the RF-PCA model based on the results of the confusion matrix. Furthermore, to verify the superiority, compared with seven different classification algorithms, the RF-PCA model achieves more reliable and accurate prediction performance in the prediction poverty levels.

The final results show that (1) The approach of feature dimension reduction not only enabled to complexity reduction of classification models, but also improves the overall accuracy of more than 3 percentage points (from 75.56% to 78.61%). (2) The RF-PCA model obtained promising results in the work of prediction poverty levels. In the future, additional data must be collected and counted to refine the present result, finding as many features as possible from SIMS.

References

[1] Zhou, Yang, Yuanzhi, Yansui, Wenxiang, Yurui: Targeted poverty alleviation and land policy innovation: Some practice and policy implications from china

[2] Luo, L.: Research on the targeted funding model of universities in the horizon of big data. Journal of Chongqing University(Social Science Edition) (2018)

[3] Shi, S.: Establishment of university subsidy system from the perspective of precision subsidy. The Theory and Practice of Innovation and Entrepreneurship (2018)

[4] Bai, J.: Analysis of student financial aid based on big data analysis. In: International Conference on Big Data Analytics for Cyber-Physical-Systems, pp. 571–577 (2019). Springer

[5] Shi, C.Y.: The enlightenment of big data thinking on the recognition of poverty-stricken students

[6] Cao, X., Wang, Y.: An approach to granting subsidies to college students in china using big data. Journal of Education and Practice **7**(26), 1–4 (2016)

[7] Xu: Identification of poor students based on campus one-card behavior data.technology and economic guide. **29**(18), 143–145 (2021)

[8] Yang: Research on the poverty level of college students' precise funding based on big data. (2018)

[9] Chao-Wen, W.U., Dai, J., Sun, Y.N.: Research on the targeted poverty reduction model of the needy undergraduates in the big data environment. Heilongjiang Researches on Higher Education (2016)

[10] Tao, B., Liu, K., Miao, F., Sun, T., Miao, R.: Targeted poverty reduction model of the needy undergraduates based on

[11] Wu, F., Zheng, Q., Tian, F., Suo, Z., Li, F.: Supporting poverty-stricken college students in smart campus. Future Generation Computer Systems **111**(1) (2019)

[12] Mohamud, J.H., Gerek, O.N.: Poverty level characterization via feature selection and machine learning. In: 2019 27th Signal Processing and Communications Applications Conference (SIU) (2019)

[13] Nguyen, C., Lo, D.: Testing proxy means tests in the field: Evidence from vietnam. Mpra Paper (2016)

[14] McBride, L., Nichols, A.: Improved poverty targeting through machine learning: An application to the usaid poverty assessment tools. Unpublished manuscript. Available at: http://www. econthatmatters. com/wp-content/uploads/2015/01/improvedtargeting_21jan2015. pdf (2015)

[15] Sani, N.S., Rahman, M.A., Bakar, A.A., Sahran, S., Sarim, H.M.: Machine learning approach for bottom 40 percent households (b40) poverty classification. Int. J. Adv. Sci. Eng. Inf. Technol **8**(4-2), 1698 (2018)

[16] Wei-jie, M.: Application of c4. 5 algorithm on determination of needy college students [j]. Journal of Henan Institute of Education (Natural Science Edition) **3** (2012)

[17] Narendranath, S., Khare, S., Gupta, D., Jyotishi, A.: Characteristics of 'escaping'and 'falling into'poverty in india: An analysis of ihds panel data using machine learning approach. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1391–1397 (2018). IEEE

[18] Irawan, S.a.M.I.: Classification of poverty levels using k -nearest neighbor and learning vector quantization methods. International jouranl of computing of science and applied mathematics **2**(1), 8–13 (2016)

[19] SHAO, W.-s., LIU, S.-d.: Application on identification of poor students based on rough set and bp neural network [j]. Coal Technology **6** (2012)

[20] Azcarraga, A., Setiono, R.: Neural network rule extraction for gaining insight into the characteristics of poverty. Neural Computing and Applications **30**(9), 2795–2806 (2018)

[21] Perez, A., Yeh, C., Azzari, G., Burke, M., Lobell, D., Ermon, S.: Poverty prediction with public landsat 7 satellite imagery and machine learning. arXiv preprint arXiv:1711.03654 (2017)

[22] Danbirni, M.I., Dongxiao, R.: Poverty prediction of nigeria by using convolutional neural network with combination of satellite image

[23] Bansal, C., Jain, A., Barwaria, P., Choudhary, A., Singh, A., Gupta, A., Seth, A.: Temporal prediction of socio-economic indicators using satellite imagery. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp. 73–81 (2020)

[24] Snelick, R., Uludag, U., Mink, A., Indovina, M., Jain, A.: Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. IEEE transactions on pattern analysis and machine intelligence **27**(3), 450–455 (2005)

[25] Ribaric, S., Fratric, I.: Experimental evaluation of matching-score normalization techniques on different multimodal biometric systems. In: MELECON 2006-2006 IEEE Mediterranean Electrotechnical Conference, pp. 498–501 (2006). IEEE

[26] Schwarz, B., Azodi, C.B., Shiu, S.-H., Bauer, P.: Putative cis-regulatory elements predict iron deficiency responses in arabidopsis roots. Plant physiology **182**(3), 1420–1439 (2020)

[27] Moisen, G.: Classification and regression trees. In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588., 582–588 (2008)

[28] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**(1), 3–42 (2006)

[29] Garreta, R., Moncecchi, G.: Learning Scikit-learn: Machine Learning in Python. Packt Publishing Ltd, ??? (2013)

[30] Louppe, G.: Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502 (2014)

[31] Wang, F., Shen, L., Zhou, H., Wang, S., Wang, X., Tao, P.: Machine learning classification model for functional binding modes of tem-1 β-lactamase. Frontiers in molecular biosciences **6**, 47 (2019)

[32] Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. In: Nips (2017)

[33] Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., Song, J.: An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. Molecular Therapy-Nucleic Acids **22**, 362–372 (2020)

[34] Molnar, C.: Interpretable Machine Learning. Lulu. com, ??? (2020)

[35] Myerson, R.: Game theory: Analysis of conflict harvard univ. Press, Cambridge **3** (1991)

[36] Mi, J.-X., Zhu, Q., Lu, J.: Principal component analysis based on block-norm minimization. Applied Intelligence **49**(6), 2169–2177 (2019)

[37] Jeffers, J.N.: Two case studies in the application of principal component analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics) **16**(3), 225–236 (1967)

[38] Chen, Q., Wynne, R., Goulding, P., Sandoz, D.: The application of principal component analysis and kernel density estimation to enhance process monitoring. Control Engineering Practice **8**(5), 531–543 (2000)

[39] Webster, T.J.: A principal component analysis of the us news & world report tier rankings of colleges and universities. Economics of Education Review **20**(3), 235–244 (2001)

[40] Varga, J.K., Tusnády, G.E.: Tmcrys: predict propensity of success for transmembrane protein crystallization. Bioinformatics **34**(18), 3126–3130 (2018)

[41] Krasoulis, A., Nazarpour, K.: Myoelectric digit action decoding with multi-output, multi-class classification: an offline analysis. Scientific reports **10**(1), 1–10 (2020)

[42] Ripoll, D.R., Chaudhury, S., Wallqvist, A.: Using the antibody-antigen binding interface to train image-based deep neural networks for antibody-epitope classification. PLoS computational biology **17**(3), 1008864 (2021)

[43] Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N.A., Trollor, J., Brodaty, H.: A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Scientific reports **10**(1), 1–10 (2020)

[44] Zhang, K., Su, H., Dou, Y.: Beyond ap: a new evaluation index for multiclass classification task accuracy. Applied Intelligence, 1–11 (2021)

[45] On machine learning methods for chinese document categorization. Applied Intelligence **18**(3), 311–322 (2003)