Hindawi
Scientific Programming
Volume 2022, Article ID 5298988, 10 pages
https://doi.org/10.1155/2022/5298988

*Research Article*

# A Statistical Analysis Model of Big Data for Precise Poverty Alleviation Based on Multisource Data Fusion

**Tian Liang and Xuefang Wang**

*Institute of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, Peking University, Beijing 100871, China*

Correspondence should be addressed to Tian Liang; liangtian2019@pku.edu.cn

Received 16 December 2021; Revised 16 January 2022; Accepted 31 January 2022; Published 14 March 2022

Academic Editor: Sheng Bin

This paper adopts the method of multisource big data fusion to conduct an in-depth study and analysis of precision poverty alleviation and uses big data statistical analysis model to explore and analyze it. Combining the characteristics of big data itself and the development of precision poverty alleviation, it focuses on the exploration of big data and introduces the background, development status, and achieved results of poverty alleviation with typical cases, followed by the analysis of the problems in the process of big data precision poverty alleviation and the study of the improvement path of big data technology precision poverty alleviation. Through the comparative analysis of the simulation accuracy of three models, the results show that the random forest model has the lowest error rate, after which the importance degree of indicators is derived using the model. In addition, the empirical analysis of the preprocessed sample data for multidimensional identification of poor households yields the contribution rate of each dimensional indicator that leads to multidimensional poverty of farm households, establishing scientific judging criteria to accurately judge whether farm households are poor on the one hand and selecting accurate identification methods to achieve accurate identification of poor households on the other hand. The tenfold crossover method is used to verify the errors in the test sample set. When the number of classification trees is greater than 100, it will gradually increase. Therefore, it is most appropriate to select the number of trees as 100. The multidimensional accurate identification model of farm household poverty constructed in this paper has an accuracy rate of 90.26% for the identification of poor households. By analyzing the accuracy rate of model identification and the contribution rate of multidimensional indicators leading to the poverty of farm households at the same time, the poverty degree of farm households under each dimensional indicator is derived, to accurately identify the poor households and their poverty status. The results show that the multidimensional accurate identification model of farm household poverty has the accurate identification ability and application value in the identification problem of poor households, and through the implementation of the model algorithm, a good application environment of accurate identification of poverty is created, which provides technical support to help poverty alleviation work and improve the accuracy of identification of poor households.

## 1. Introduction

Modern technology is increasingly linked with government management, and management through modern technology means not only enables citizens to obtain more convenient government services but also carries the innovative initiatives and determination of the state to create modern management means. What the impact of the development of big data technology on the government's precise poverty alleviation is and how to use big data technology to achieve modern scientific and effective precise poverty alleviation management under this impact are worthy of in-depth consideration and research. Aware of the importance of using big data technology in management, we can promote the progress and development of management information technology, including big data technology. Big data technology has become an indispensable tool in modernized precision poverty alleviation, which can effectively improve the efficiency of poverty alleviation and enhance the quality of government governance. Faulty damage is a common occurrence. However, the economic strength of poverty-stricken areas itself is backward, and the proportion of resources that can be allocated to vocational education is very small. The core of intelligent analysis research on precise poverty eradication contains the prediction of the time to get out of poverty and the generalization of the rules of help measures [1]. The essence of time out of poverty prediction and implementation rules for helping measures is to dig deep into the relationship between poor households, helping measures, and poverty alleviation based on existing poverty alleviation data, the former realizes the mathematical quantification of the inner law between "poor households - helping measures poverty" and explores the mechanism between poor households' characteristics and helping measures [2]. The latter clarifies the principle of the rules between the basic information of poor households and help measures and further clarifies the correspondence between the characteristics of poor households and help measures. The purpose of the research on the intelligent analysis of precise poverty eradication is to use the generated rule set for the implementation of help measures to formulate a help plan for poor households, and then evaluate and adjust the poverty eradication plan by predicting the time of poverty eradication, to finally achieve the maximum utilization of resources and the fastest and most stable poverty eradication of poor households [3].

In the early stage of poverty alleviation work, due to the large amount of poverty alleviation data and because the traditional way of storing farm household information is mostly based on paper materials or spreadsheets, there are problems such as inaccurate and nontransparent, incomplete, or untrue information collection, resulting in relatively disorganized storage of farm household files, leading to easy falsification of poverty alleviation object information data, making poverty alleviation information incomplete and difficult to retrieve [4]. The backwardness of data collection tools, coupled with the untimely update of poverty information and the lack of dynamic management, makes it more difficult for poverty alleviation departments at all levels to accurately determine the real situation of farm households; there is also a deficiency in the application services of poverty alleviation data, which cannot meet the needs for rapid information search, data mining, statistical analysis, etc. and cannot make accurate judgments on incomplete data in all aspects based on the audit results. It has the advantages of good stability and flexibility. The model is applied to the precision poverty alleviation data analysis system and achieves good results. Therefore, whether in the issue of identification of poor households or the direction of identification of causes of poverty, there is a lack of more accurate means of identification, making the final determination of poor household candidates inaccurate, and in the process of research and formulation of the poverty criteria system, it fails to take into account the indicators affecting the living standards comprehensively, including multidimensional indicators such as the number of family members in the labor force, education level, housing situation, and the policies and benefits enjoyed. This may also result in a lack of precision in determining whether a farm household is poor [5].

In the process of implementing and promoting the actual education subsidy work, there are still problems such as backward means of subsidy, incomplete and inaccurate information of students, poor information among related departments, and untimely and inaccurate subsidy for students, which seriously restrict the efficiency and accuracy of education subsidy work. This research focuses on how to use big data analysis technology in education financial aid work to realize the analysis of financial aid index for students from poor families with established records and cards, accurate identification and pushing of poor students based on integrating students' information in various aspects, providing decision support for education financial aid work through education financial aid index analysis, and improving the efficiency and accuracy of education financial aid. At the same time, an education precision poverty alleviation system is established based on information integration to improve the level of informatization of education subsidy work to promote better education subsidy work.

## 2. Related Work

It is believed that big data can support governments in moving towards better policy and public management goals and contribute to more effective management and policy analysis, thereby facilitating resource allocation. Big data will also continue to improve the management of public programs at all levels of government, thus contributing to the development of efficient and innovative government. It is believed that big data plays a positive role in government management, predicting many valuable policy outcomes, while big data will improve our description of most public policy issues [6]. The degree of benefit derived from big data varies for different sectors and difficulties faced in applying big data, while government faces the least difficulties, gains more, and has greater value potential in applying big data [7]. The report also suggests that the effective use of big data can create tremendous value by using it to improve resource allocation and coordination, reduce waste, enhance transparency, and facilitate the generation of new ideas and insights. The theory lays the theoretical foundation for big data technology for precise poverty reduction [8].

In addition, by focusing on the background of big data, the transformation of national governance has also been studied by many scholars, which provides good guidance for big data to help government work. For example, the "book smart government" tells some innovative ideas of big data in the field of government governance, which believes that big data is not only a kind of massive data state and processing technology, but also a way of thinking; proposes to introduce the means and methods of big data into the management field, which can realize the modernization of management; and advocates that the government can tap the huge value from the huge amount of data and become the leader in the era of big data [9]. The difficulties of implementing a big data governance approach in government are analyzed, and the significance and impact of big data in terms of improving the governance capacity of our government are analyzed, effectively giving guidance on government work [10]. A data management platform for precise poverty alleviation is designed and implemented, and a sound and complete file of poor households is established. For the first time, the construction of a big data platform for precise poverty alleviation was proposed to centralize the management of poverty alleviation targets, help measures, poverty alleviation effectiveness, and performance assessment. However, the information management for poverty alleviation data is not deep enough, so it is proposed to use information technology to realize the intelligent analysis of poverty alleviation data [11].

Most of the current research focuses on the accurate identification of poor households with the help of statistical analysis methods to accurately identify poor households from multiple dimensions. AdaBoost method in machine learning is used to construct a poor household identification model. The multidimensional fuzzy poverty index is constructed based on the fuzzy set method for accurate identification and classification of poor people and their poverty level [12]. The random forest model in data mining is used to analyze the characteristic items of poor households to select the main factors that influence the poor households to get out of poverty. The correlation analysis of poverty household identification calibration rules based on poverty data from ethnic minority areas is carried out using data analysis methods, and the idea of implementing a poverty dynamic evaluation index model is proposed. The work only stays in the information management of poverty alleviation and shallow data analysis, lacking accurate portrayal and quantitative analysis of poverty alleviation effectiveness and time out of poverty, and the exploration and research on the implementation rules of help measures are not deep enough [13]. The main reasons for this are, on the one hand, the unique social form of China, the complexity of the poverty situation, the wide range of areas involved, and the difficulty of in-depth integration of data communication and exchange among various departments; on the other hand, there is the considerable challenge of constructing a suitable mathematical model to accurately describe these complex factors.

## 3. Analysis of a Statistical Analysis Model of Precision Poverty Alleviation Big Data with Multisource Data Fusion

*3.1. Multisource Data Fusion Design for Big Data.* To ensure that the interference of data noise, missing values, and inconsistent data is received during the modeling process, data cleaning transformations are needed to ensure that the data is accurate, complete, and consistent. Among them, data cleaning includes the processing of missing values, outliers, and incorrect values, and data transformation includes the processing of character variables. Several operations of data cleaning changes covered in this paper are described in detail below [14]. However, some errors are difficult to correct, especially for time-sensitive and scenario-specific data. Objects lead to errors in the data on household income. In such cases, missing values can be substituted or, if the error item is a required item, only deletion measures can be taken. The acceptance rate calculated by the two methods is gradually decreasing, and on the right side of the dotted line, the two test methods reject the null hypothesis with a trend close to 1.

In the traditional poverty alleviation process, the focus is on the income and livelihood of poor households, while the causes of poverty, external factors, and internal conditions of households are neglected, making it possible to invest many resources in poverty alleviation with less than satisfactory results. The reason is the failure to consider and calculate multiple deep-rooted poverty-causing factors in an integrated and comprehensive manner. The causes of poverty vary among the poor, as do their economic bases and the requirements of the assistance recipients, and the complexity of the situation makes the precise identification of the main causes of poverty biased.

$$\theta = \frac{m_+ - \left(m_1 + m_-\right)}{m_+ + m_-}, \tag{1}$$

where$m_+$ denote the number of positive and negative cases in the training dataset, respectively, and $m_-$ are the number of positive and negative cases covered by the rule set. The distance between samples is a measure of similarity between two samples, which can be calculated directly by the distance formula; the distance between samples and clusters is a measure of similarity between samples and elements in clusters, which usually calculates the centroids of clusters first and uses the distance between samples and cluster centroids to represent the distance between samples and clusters; the distance between clusters is a measure of similarity between clusters and clusters, which usually calculates the centroids of each cluster separately first and uses the distance between cluster centers [15]. The common distance algorithms used in K-Means clustering algorithm are Euclidean distance, Manhattan distance, and Minkowski distance, as shown in Figure 1. The distribution of family members is concentrated in 3–5 people, and the per capita annual income of the family is 5,000 to 17,000 yuan. The characteristics of the third category of poor students are that the education level is junior high school, the overall health status of family members is healthy, the distribution of family members is concentrated in 5–6 people, and the per capita annual income of the family is 17,000 to 35,000 yuan.

For the current dataset, the information gain is calculated separately for each attribute.

$$\begin{aligned}
\text{Entropy}(S) &= \sum_{i=1}^{m} p_i \quad \ln\left(p_i^2\right),\\
\text{Entropy}_A(S) &= \sum_{j=1}^{m} \frac{S_j}{S} p_i \quad \ln\left(p_j^2\right),\\
\text{Gain}_A(S) &= \text{Entropy}(S) + \text{Entropy}_A(S).
\end{aligned} \tag{2}$$

In calculating the information gain city of an attribute, first calculate the information entropy of the dataset $\text{Entropy}(S)$, and the$\text{Entropy}_A(S)$, $\text{Entropy}(S)$ the

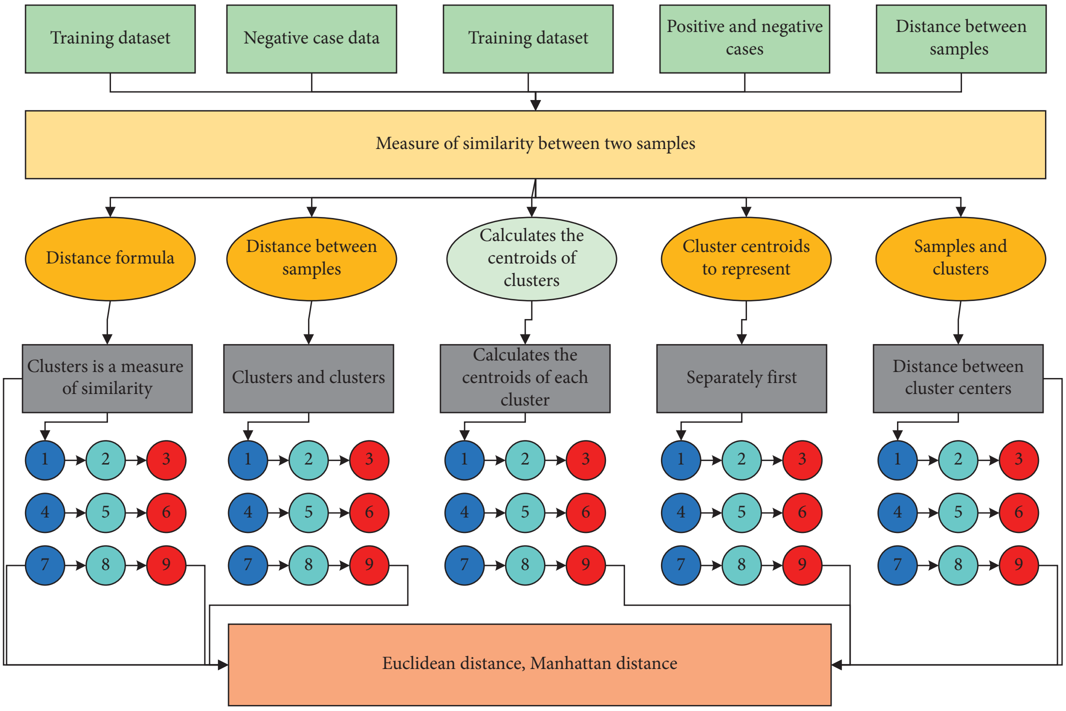
Training dataset
Negative case data
Training dataset
Positive and negative cases
Distance between samples
Measure of similarity between two samples
Distance formula
Distance between samples
Calculates the centroids of clusters
Cluster centroids to represent
Samples and clusters
Clusters is a measure of similarity
Clusters and clusters
Calculates the centroids of each cluster
Separately first
Distance between cluster centers
1
2
3
4
5
6
7
8
9
Euclidean distance, Manhattan distance

Figure 1: Multisource data fusion framework.

$\text{Entropy}_A(S)$ difference between and is the information gain $\text{Gain}_A(S)$. Obviously $\text{Entropy}_A(S)$ the smaller and the $\text{Gain}_A(S)$ larger, the greater, indicating that attribute A greatly reduces the information entropy required to classify the dataset S. The algorithm is recursively invoked for the subdataset, and if the category attribute of the subdataset contains only a single attribute, it means that the branch has finished splitting, the split attribute is then the leaf node, and then it returns to the invocation. The core of performing network parameter updates is to use the mean square error as the objective function and a gradient descent strategy to adjust the parameters along the negative gradient direction of the objective function, as exemplified by the connection right $w_{hi}$ from the hidden layer to the output layer.

$$\Delta w_{hi} = \eta \frac{\partial E_k}{\partial w_{hi}}, \tag{3}$$

$$\frac{\partial E_k}{\partial w_{hi}} = \frac{\partial E_k}{\partial y_j^k} \cdot \frac{\partial y_j^k}{\partial \beta} \cdot \frac{\partial \beta}{\partial w_{hi}}, \tag{4}$$

where $\eta$ is the given learning rate and takes values in the range (0, 1). Substituting the expressions of $E_k$ and $\beta$ into the partial derivative equation (4) yields $g_j$ and $b_h$ as shown in (5). The standard BP algorithm, although powerful, also has some drawbacks. Since the BP algorithm minimizes the error $E$ between the actual and output values by changing the parameters, the gradient-based search method may cause the model to fall into the trap of local minima and global minima, which will directly affect the accuracy of the model fit. On the other hand, the convergence time of the model is also an important factor to be considered in the modeling process [16]. In addition, a more in-depth analysis of the above two problems is carried out, and a model framework is proposed. Combined with the specific problems of the model realization, the variable screening and preprocessing work is done on the obtained data. Ideally, the optimal parameters need to be obtained using the minimum convergence time. In the standard BP algorithm, the initial parameters of the model are obtained randomly, and there is uncertainty in the parameter selection, so their selection also has an impact on the accuracy and convergence time of the model to some extent, and suitable initial parameters may make the model reach the convergence point faster and more accurately. Some scholars propose to use swarm intelligence optimization algorithms such as genetic algorithms and particle swarm algorithms to improve the parameter selection or update of the BP algorithm.

$$g_j = -\frac{\partial E_k}{\partial y_j^k} \cdot \frac{\partial y_j^k}{\partial \beta} \cdot \frac{\partial \beta}{\partial w_{hi}}, \tag{5}$$

$$b_h = -\frac{\partial \beta}{\partial w_{hi}}. \tag{6}$$

The correlation analysis method using redundancy measure between features is used for feature selection. The main idea of this method is to measure the redundancy between attributes by measuring the correlation between them. The advantage of the above algorithm is that by using the filter relief algorithm with high computational efficiency and no restrictions on the size and type of datawe can find out those features that are not relevant to the target attributes, and then combine with the hierarchical and relevance analysis to solve the problem together, as shown in Figure 2.

The hypothesis testing problem for the multisource data fusion problem has been introduced; in this problem, we want to test whether a particular node with the help of data from other nodes can improve the coefficient estimates of its nodes, i.e., whether it will reduce the mean squared error (MSE) of the coefficient estimates of a single node. The hypothesis problem we are interested in is

$$\left\|G^{-2}\Delta\beta\right\|_2^2 \geq \sigma_1^2. \tag{7}$$

If $G$, $\Delta\beta\sigma_1^2$ is known as a natural test statistic of

$$U\left(G^{-2}, \Delta\beta, \sigma_1^2\right) = \left\|\frac{G^{-2}\Delta\beta}{\sigma_1}\right\| + \lambda\left\|\frac{G^{-2}\cdot\Delta\beta}{\sigma_2^3}\right\|, \tag{8}$$

where $\sigma_1^2$ is the variance of the noise at node. To explore the importance measure of indicators and at the same time provide a basis for the subsequent application of the two-level weight assignment method combining entropy weight method and prior knowledge to finally determine the weight of each indicator, this paper implements three models based on language. Big data technology has become an indispensable tool in modern precision poverty alleviation, which can effectively improve the efficiency of poverty alleviation and improve the quality of government governance.

The standard BP algorithm is powerful but has some drawbacks. Since the BP algorithm minimizes the error $E$ between the actual value and the output value by changing the parameters, the gradient-based search method may cause the model to fall into the trap of local minima and global minima, which will directly affect the accuracy of the model fit; on the other hand, the convergence time of the model is also an important factor to be considered in the modeling process [17]. Ideally, the minimum convergence time needs to be used to obtain the optimal parameters. In the standard BP algorithm, the initial parameters of the model are obtained randomly, and the parameter selection has uncertainty, so its selection also has an impact on the accuracy and convergence time of the model to some extent, and the appropriate initial parameters may make the model reach the convergence point faster and more accurately. Some scholars propose to use swarm intelligence optimization algorithms such as genetic algorithms and particle swarm algorithms to improve the parameter selection or update of the BP algorithm.

*3.2. Design of Statistical Analysis Model for Accurate Poverty Alleviation Big Data.* The information entry and maintenance module are detailed into the village collective questionnaire part and the household research questionnaire part [18]. The household research questionnaire section contains basic information about the poverty alleviation targets, including their suggestions to the village committee and government as well as household farming and animal husbandry. There is a one-to-many relationship of a household with multiple farming and livestock farming, and suggestions to the village committee and government can be added in multiple entries. Then, the poverty alleviation plan is evaluated and adjusted through the forecast of poverty alleviation time, and finally the maximum utilization of resources and the fastest and most stable poverty alleviation of poor households can be achieved.

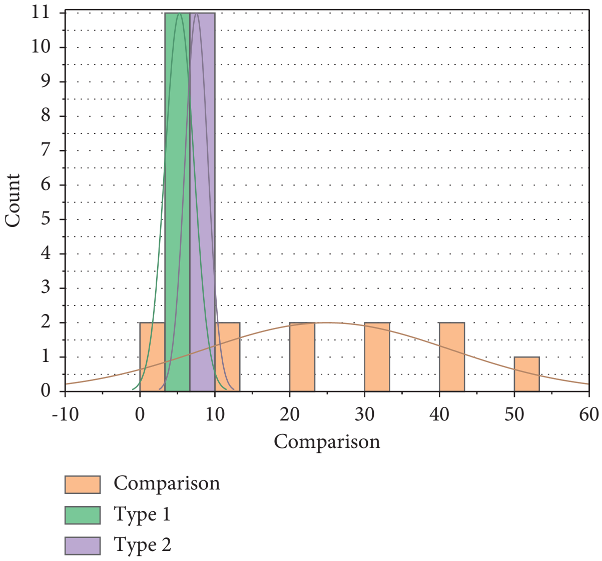

Figure 2: Frequency histogram comparison of the values taken for the test statistic.

The random forest model with a relatively small error rate is obtained by using a combination of simulation and a tenfold crossover validation test. The error in the test sample set using the tenfold crossover method tends to increase when the number of classification trees is greater than 100, so 100 trees are most appropriate. The average error rate in the test sample set is 16.74% compared to the average error rate of 0 in the training sample set. Therefore, the overall recognition accuracy of the model is 83.26%, as shown in Figure 3.

Information technology in poor areas has a late start, a low level, and a poor foundation, and the construction of facilities mostly remained at the level of developed areas decades ago [19]. If we want to play the role of information technology in vocational education for precise poverty alleviation, we must first invest a lot of money in information technology infrastructure, including information technology hardware and software equipment, network bandwidth, and information resource base. There are problems such as inaccurate, opaque, incomplete, or untrue information collected, resulting in relatively cluttered storage of farmers' archives, which makes the information and data of poverty alleviation objects easy to falsify. Vocational education, compared with general education, has higher requirements for sites, equipment, and talents for information technology equipment. In addition to the basic equipment of facilities, specialized talents are needed to participate in maintenance and management. Information technology equipment is a consumable item, and failure damage is common. However, the poor areas themselves are economically backward, and

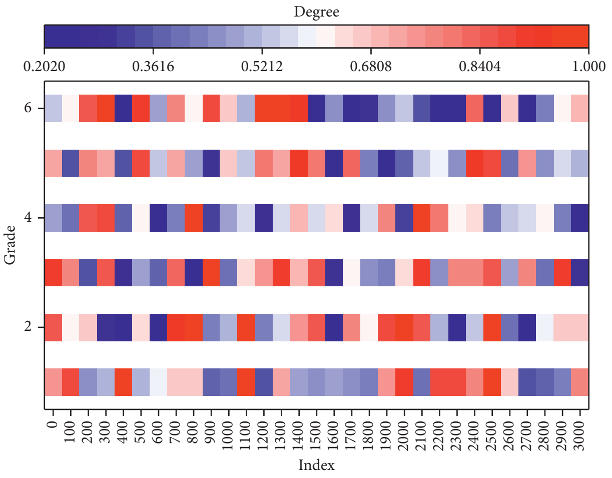

Figure 3: Distribution of model error rates.

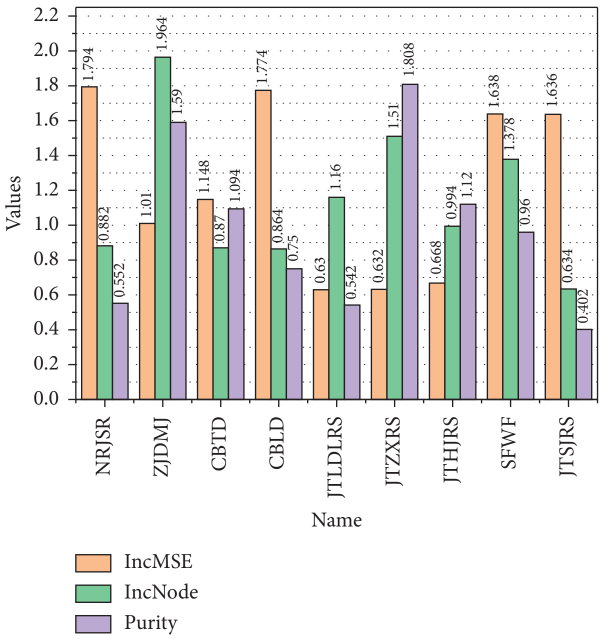

Figure 4: Level of importance of the characteristic variables.

the proportion of resources that can be allocated to vocational education is even more insignificant.

To obtain a model with higher accuracy, facilitate the use of the model to determine the importance of indicators, and ultimately make an auxiliary reference for determining the weights of each indicator, it is clear from the comparison of experimental results in this paper that the RF model has higher accuracy. Therefore, by using the importance function in the RF model algorithm, the importance values of the indicators included in the model fitted by random forest can be derived, and the degree of importance of the characteristic variables (indicators) is shown in Figure 4. In the process of researching and formulating the poverty standard system, the indicators affecting the living standards were not fully considered, including the number of family members in the labor force, education level, housing conditions, and enjoyment. Multidimensional indicators such as policy and welfare also make it inaccurate to determine whether a farmer is poor or not. The results show that the top four characteristic variables (indicators) in terms of the importance of impact when identifying poor households from multidimensional dimensions are the area of the house base (ZJDMJ), annual per capita income (NRJSR), contracted land (CBLD), and whether the house is in danger (SFWF). In this paper, based on the multidimensional poverty index (MPI), a global dimensional indicator system proposed by the Human Development Index (HDI) developed by UNDP, the indicator of the home base area, which has an important measure of 85.82%, and the indicator of annual per capita income, which has an important measure of 81.23%, are used as the main basis for determining whether a farm household is poor, as shown in Figure 4.

Firstly, a series of preprocessing methods were applied to the source dataset for data preparation, and then a variety of methods were used in feature selection for comparison experiments. Furthermore, since it was considered that the information data based on multidimensional poverty contains both discrete and continuous values, and the feature data of the series was characterized by hierarchical nature, a novel feature selection algorithm based on REAHCORalgorithm was used in feature selection, applied to GBDT classifier with high classification accuracy, and achieved an evaluation model with better classification results. The innovative proposed REAHCOR algorithm can not only reduce the dimensionality of the huge data feature set, but also ensure the strong classification ability of the features after the reduction, and the evaluation effect of the overall model is also verified, which has the advantages of good stability and flexibility [20]. Once this situation is found, it can be deleted, filled with values, or assigned as missing values. For example, an obvious error occurred in the registration of the area of land owned by poor households: the arable area of 5 mu of households was regarded as the area of arable land per capita. The model is applied to the precise poverty alleviation data analysis system to achieve good results.

The implementation of the accurate poverty alleviation data analysis system consists of entity layer, DAO layer, service layer, and controller layer. Each layer implements specific functions, and in the case of poverty level evaluation, the DAO layer provides many unified interfaces, such as “getData” as the interface to get data, “save” as the interface to save data, “delete” as the interface to delete the current record, and “use the form” to submit a form to be saved to the database.

It implements the entry of questionnaire forms, including basic village information as well as information on villagers belonging to the village. It covers the economic and cultural development of the village and the village cooperative enterprises. Among them, the village economic and cultural development situation is entered into the form through the village name link, which is filled in by month. The village coorganized enterprises belong to the village, and there is a relationship that one village corresponds to multiple coorganized enterprises. The household research questionnaire section contains basic information about the poverty alleviation targets, including their suggestions to the village committee and government as well as household farming and animal husbandry. There is a relationship between the household research and the village collective information to which they belong. There is a one-to-many relationship of a household with multiple farming and livestock farming, and suggestions to the village committee and government can be added in multiple entries.

## 4. Analysis of Results

*4.1. Data Fusion Results.* Since the simulation algorithm for the test utilizes only the approximate chi-square distribution, only the outer loop is required to calculate the Type I error probabilities. In contrast, the PB test requires an inner loop to generate a self-help sample from the estimated model and calculate the $p$ value using the Monte Carlo method; the outer loop generates a sample of observations from the set-up parametric model and calculates the acceptance rate for the above hypothesis test. It is difficult to select the absolute randomness of the sample. If any aspect of random data is ignored, it is difficult to form an accurate analysis.

Figure 5 shows a line graph comparing the acceptance rates calculated by the methods used in the parametric bootstrap approach proposed in this section, sharing all coefficients. The red line shows the acceptance rates calculated using the parametric bootstrap approach, while the blue line shows the acceptance rates calculated using the methods in the literature. The solid purple line shows the value of $n$ when the sufficient condition in theorem is equal to 1. The dashed blue line shows the value of $n$ when the MSEs of the two models are equal. On the right side of the dashed line, the MSE of the single-node model is smaller than that of the two-node model. On the left side of the dashed line, the situation is exactly the opposite. Figure 5 shows that in the left part of the solid blue line, the probability of the parametric bootstrap method committing the first type of error is small and stable around 0.05, while the probability of the first type of error for the “literature” test is large and the worst at the sample size of $n = 2^4$. As the sample size gradually increases, the acceptance rate calculated by both methods gradually decreases, and on the right side of the dashed line, the two tests reject the original hypothesis with a potential close to 1. As seen in Figure 5, the parametric bootstrap test proposed in this section can better control the first type of error probability, has no less potential than the existing test, and is more meaningful in practical applications. The gradient-based search method may make the model fall into the trap of local minimum and global minimum, which will directly affect the accuracy of model fitting; on the other hand, the convergence time of the model is also an important factor to be considered in the modeling process.

The classical gradient descent optimizer is chosen as the optimizer in the backpropagation optimization method with the learning rate parameter; moreover, to combine the advantages and disadvantages of the descent algorithm and the stochastic gradient descent algorithm, the experiments compute the loss function for a small portion of the training data at a time, which is called a batch. By matrix operations, the parameters of the neural network are optimized in batches and will not be much slower than the individual data. The learning rate is set using exponential decay, and L2 regularization is used to avoid the overfitting problem and limit the weight size so that the model cannot arbitrarily fit the random noise in the training data.

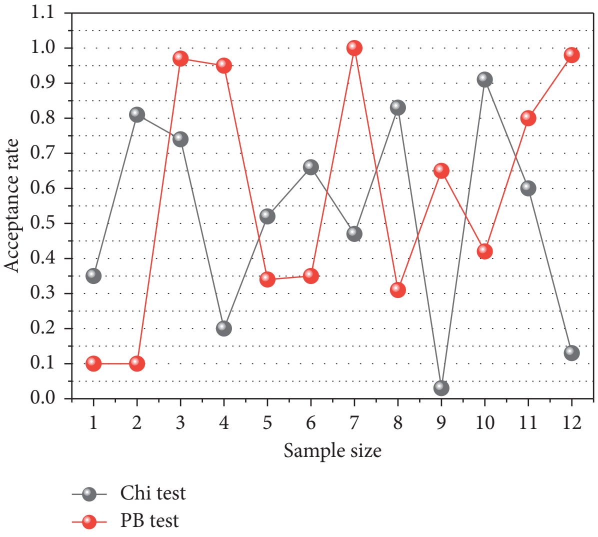

Figure 5: Comparative line graph of acceptance rates when sharing all coefficients.

After completing the training of BPNN, FOA-BPNN, and DSFOA-BPNN with the training set, the three types of models were tested on the same test set, and the prediction accuracy rate and loss function change curves of each model were obtained as shown in Figure 6. From Figure 6, the trend of the change of the accuracy rate of the out of poverty time prediction models constructed using the three methods is as follows: at the early stage of training, the accuracy rate of the models changes rapidly; as the training proceeds, the accuracy rate decreases gradually , and the models tend to stabilize; and all three methods can successfully fit the out of poverty time problem without considering the high or low accuracy rate.

The initial prediction accuracies of the three types of models differed significantly, with the initial prediction accuracies of the BPNN, FOA-BPNN, and DSFOA-BPNN models being 0.28, 0.42, and 0.49, respectively, and the model accuracies under the initial parameters of DSFOA-BPNN being 0.21 and 0.07 higher than those of BPNN and FOA-BPNN, respectively; the fundamental reason for these results is that BPNN was randomly selected using a normal distribution probability model, FOA-BPNN was filtered by a standard Drosophila optimization algorithm, and DSFOA-BPNN was selected by a modified dynamic Drosophila optimization algorithm on the initial parameters. Thus, the initial parameters that have been merited by the FOA or DSFOA algorithms are significantly more accurate than the randomly selected ones making the initial prediction of the model. Use the correlation analysis method to measure the

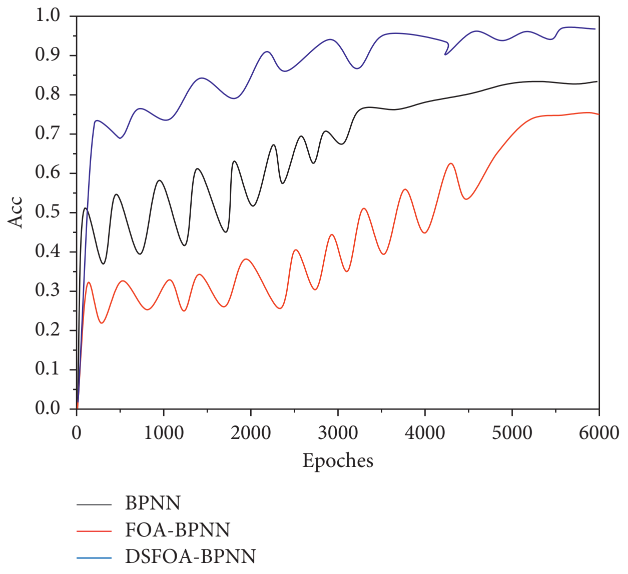

Figure 6: Change inaccuracy of model time out of poverty prediction.

redundancy of the paired features in the set U, delete the feature with a smaller weight in the set S between the two features whose redundancy is greater than the redundancy threshold in the result set, and select the final the required subset of features; these selected features are some features that are strongly related to the category label.

The prediction accuracy of the DSFOA-BPNN model is stable at 0.52, 0.62, and 0.70. By analyzing the three results, FOA-BPNN improved by 0.10 compared to the original BPNN model, while DSFOA-BPNN improved by 0.08 compared to FOA-BPNN. The fundamental reason for the above results lies in the initial parameter selection. The initial parameters not only determine the initial accuracy of the model but also affect the ability of the model to jump out of the local optimum, to some extent.

In turn, it affects the final prediction accuracy of the model. The above experimental results show that the initial parameter selection of FOA and DSFOA for BP neural network can avoid the dilemma of BPNN falling into local extremes to a certain extent, thus improving the prediction accuracy; on the other hand, the dynamic fruit fly optimization algorithm (DSFOA) optimizes the initial parameters better than the standard fruit fly optimization algorithm, and the flexible step size variation of the former makes the results of iterative optimization search more accurate.

*4.2. Results of the Analysis of Precision Poverty Alleviation.* The integrated dataset has more comprehensive student information, which includes students' basic personal information, family information, and family economic situation, but not all data need to be analyzed, such as students' names, ID card numbers, and home addresses. Therefore, it is necessary to filter out the needed data after the data preprocessing is completed and filter out the data not needed for modeling. In this paper, we mainly study the use of a decision tree algorithm to establish the poor family identification model. The information related to a family economic situation such as labor skills, a new agricultural cooperation, health status, and net income per capita in student information can best reflect the poverty situation of students' families, so we select seven data fields related to poverty identification from the integrated dataset of student information of poor families after the model requirements, which are labor skills, a new agricultural cooperative, health status, rural pension insurance, and poverty identification. The seven data fields related to poverty identification are labor skills, new agricultural cooperation, health status, rural pension insurance, family size, per capita net income, and low-income households.

From the decision tree constructed for poverty identification, we can see that the identification of students' poverty status is mainly based on the per capita annual income of poor students' families and whether the students' families are poor or not. The model has a high degree of interpretability and is consistent with the actual poverty identification work.

In the actual education subsidy work, due to the dynamic nature of the existing poor students' data on the one hand and because new poor students' data are added to the database every year on the other hand, the poverty status identification model trained in this paper is used to accurately identify the poor students' poverty status according to the dynamically changing poor students' data, and the "poverty status identification" (reg_poor_status) attribute is added to the students' data. The result after adding the "poverty status recognition" attribute is shown in Figure 7, where the class is the current poverty status of poor students and reg_poor_status is the result of recognition based on the poverty information of poor students using the poverty recognition model built in this paper. The comparative analysis of class and reg_poor_status can help the education subsidy department to identify students who are currently enjoying the subsidy but may have reached the poverty standard and students who are not currently poor but may meet the poverty subsidy standard from the list of poor students, to improve the accuracy of the financial aid work for poor students.

From the probability distribution chart, the first category of poor students has the following characteristics: the average education level of family members is high school, the overall health condition of family members is healthy, the distribution of family size is concentrated in 4–7 persons, and the annual income per capita of the family is 35,000 to 60,000 yuan. The characteristics of the second category of poor students are as follows: the average education level of family members is primary school, the overall health condition of family members is a chronic illness, the distribution of family members is concentrated in 3–5 persons, and the annual per capita income of the family is 5,000 to 17,000 yuan. The third category of poor students is characterized by the following: the evaluation of the education level is junior high school, the overall health condition of family members is healthy, the distribution of family members is concentrated in 5–6 persons, and the annual per capita income of the family is 17,000 to 35,000 yuan, which shows that the

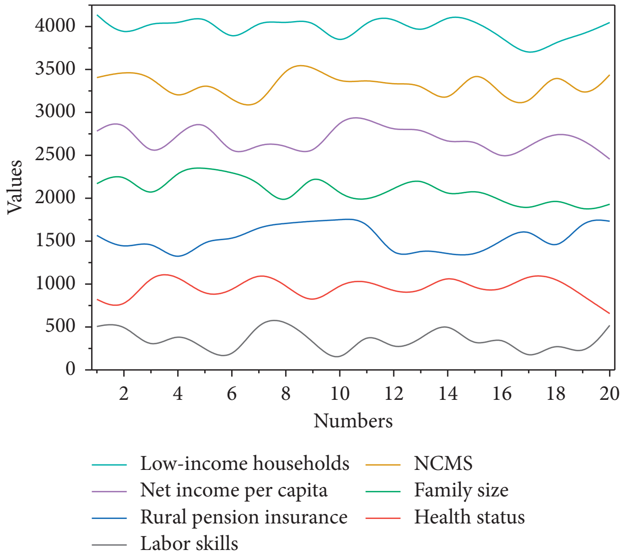

Figure 7: Poverty status recognition model result graph.

second category of poor students has a smaller number of family members and a shorter labor force, the overall average education level of family members is low, and the annual per capita income is low. Appropriate initial parameters may make the model reach the convergence point faster and more accurately. The percentage of poor students in this category is 83.11%, and the situation of these poor students is more difficult than that of poor students in the first and third categories, so they should be the priority target of educational subsidies.

At the same time, the overall average literacy level of the poor families in the file card is low. The implementation of education subsidies to promote educational equity and improve the quality of education and education of poor students is an important aspect of the implementation of the work of precise poverty alleviation. The clustering results provide a certain reference basis for the work of educational subsidies and help promote the work of precise poverty alleviation in education more precisely, as shown in Figure 8.

For the help measure rule induction problem, a detailed analysis of the module framework diagram of the help measure rule induction problem is presented, and a solution for rule induction using the RIPPER algorithm in rule learning is proposed; based on the student loan disbursement data, the coverage accuracy of the rule sets generated by C4.5, PART decision tree, and RIPPER is compared on the test dataset, and the experimental analysis proves that the proposed rule induction problem using the rule learning algorithm RIPPER is effective for the help measure rule induction problem.

From the theoretical level, it discusses the theoretical and informational development of precise poverty alleviation work and lists the type achievements between them; it explains the shortcomings of precise poverty alleviation informatization and further discusses the meaning and research significance of precise poverty eradication intelligent analysis; it specifies the overall framework of precise poverty eradication intelligent analysis research and analyzes the relationship between "measurement rules" and "induction." The overall framework of the research on precise poverty eradication intelligent analysis is elaborated, the connection between "measure rule summarization" and "time prediction of poverty eradication" and their positions in the work of precise poverty eradication are analyzed, a more in-depth analysis of the above two problems is carried out, and the model framework is proposed. Combined with the specific problems of model implementation, variable screening and preprocessing of the obtained data are completed.

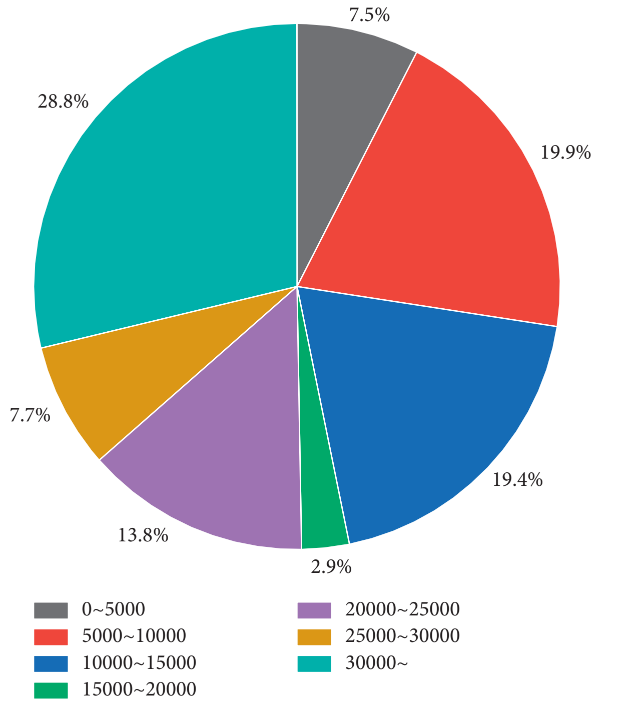

Figure 8: Distribution of annual per capita income of poor students' families.

## 5. Conclusion

By analyzing the example of applying big data to the industry, to grasp more clearly the direction of the research of the precise poverty alleviation data analysis system, the beginning section of this paper elaborates on the research background and research significance of the system, which can be developed to better assist the poverty alleviation team to accurately identify poor households, accurately help the poor people, and improve the efficiency of poverty alleviation work. Then, the status of research at home and abroad is analyzed, including the development status of precise poverty alleviation information management and the research status of poverty classification and prediction, to analyze the project and deepen the understanding of poverty-related content, which helps the subsequent sections more comprehensively. A more important research component is the in-depth study and thorough proficiency in the relevant technologies needed to implement the system. The special feature of this paper is the establishment of a poverty ranking evaluation model, which is done through the data collected from the farmers' information in the previous stage and based on data mining techniques. Therefore, this paper provides a brief introduction to data mining techniques, summarizes the current development status of data mining by reviewing a large amount of literature, and examines the main analytical methods used in data mining techniques. The poverty level evaluation model in this paper is characterized as a classification model, so this paper has studied and researched the common classification algorithms in depth, with profound understanding of their principles and usage.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Institute of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, Peking University.

## References

[1] A. Huang and F. Wu, "Two-stage adaptive integration of multi-source heterogeneous data based on an improved random subspace and prediction of default risk of micro-credit," *Neural Computing & Applications*, vol. 33, no. 9, pp. 4065–4075, 2021.

[2] X. Chen, H. H. Wang, and B. Tian, "Visualization model of big data based on self-organizing feature map neural network and graphic theory for smart cities," *Cluster Computing*, vol. 22, no. 6, Article ID 13305, 2019.

[3] B. Chen, B. Xu, and P. Gong, "Mapping essential urban land use categories (EULUC) using geospatial big data: progress, challenges, and opportunities," *Big Earth Data*, vol. 5, no. 3, pp. 410–441, 2021.

[4] Y. Yao, J. Zhang, Y. Hong, H. Liang, and J. He, "Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data," *Transactions in GIS*, vol. 22, no. 2, pp. 561–581, 2018.

[5] B. Majeed, J. Peng, A. Li, R. Delgadob, and Y. Lin, "Forecasting the demand of mobile clinic services at vulnerable communities based on integrated multi-source data," *IISE Transactions on Healthcare Systems Engineering*, vol. 11, no. 2, pp. 113–127, 2021.

[6] J. Han, Z. Zhang, Y. Luo et al., "The RapeseedMap10 database: annual maps of rapeseed at a spatial resolution of 10 m based on multi-source data," *Earth System Science Data*, vol. 13, no. 6, pp. 2857–2874, 2021.

[7] A. M. Snauffer, W. W. Hsieh, A. J. Cannon, and M. A. Schnorbus, "Improving gridded snow water equivalent products in British Columbia, Canada: multi-source data fusion by neural network models," *The Cryosphere*, vol. 12, no. 3, pp. 891–905, 2018.

[8] B. Wu, F. Tian, M. Zhang, H. Zeng, and Y. Zeng, "Cloud services with big data provide a solution for monitoring and tracking sustainable development goals," *Geography and Sustainability*, vol. 1, no. 1, pp. 25–32, 2020.

[9] R. Nie, J. Cao, D. Zhou, and W. Qian, "Multi-source information exchange encoding with pcnn for medical image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 986–1000, 2020.

[10] Y. Guo, X. Hu, Z. Wang et al., "The butterfly effect in the price of agricultural products: a multidimensional spatial-temporal association mining," *Agricultural Economics*, vol. 67, no. 11, pp. 457–467, 2021.

[11] C. Deng and W. Lin, "Day and night synergy to improve subpixel urban impervious surface mapping in desert environments at 30-m Landsat resolution," *International Journal of Remote Sensing*, vol. 41, no. 24, pp. 9588–9605, 2020.

[12] R. Patidar, S. M. Pingale, and D. Khare, "An integration of geospatial and machine learning techniques for mapping groundwater potential: a case study of the Shipra river basin, India," *Arabian Journal of Geosciences*, vol. 14, no. 16, pp. 1–16, 2021.

[13] J. Madhuri and M. Indiramma, "Role of big data in agriculture," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 2, pp. 3811–3021, 2019.

[14] Y. Hua, X. Wang, Y. Li, P. Xu, and W. Xia, "Dynamic development of landslide susceptibility based on slope unit and deep neural networks," *Landslides*, vol. 18, no. 1, pp. 281–302, 2021.

[15] Y. Liu, X. Ma, L. Shu, G. P. Hancke, and A. Mahfouz, "From Industry 4.0 to Agriculture 4.0: current status, enabling technologies, and research challenges," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4322–4334, 2020.

[16] R. Sun, G. Gao, Z. Gong, and J. Wu, "A review of risk analysis methods for natural disasters," *Natural Hazards*, vol. 100, no. 2, pp. 571–593, 2020.

[17] Q. Cheng, R. Oberhänsli, and M. Zhao, "A new international initiative for facilitating data-driven Earth science transformation," *Geological Society, London, Special Publications*, vol. 499, no. 1, pp. 225–240, 2020.

[18] Y. Zhou, Z. Miao, and F. Urban, "China's leadership in the hydropower sector: identifying green windows of opportunity for technological catch-up," *Industrial and Corporate Change*, vol. 29, no. 5, pp. 1319–1343, 2020.

[19] W. J. Niu, Z. K. Feng, W. F. Yang, and J. Zhang, "Short-term streamflow time series prediction model by machine learning tool based on data preprocessing technique and swarm intelligence algorithm," *Hydrological Sciences Journal*, vol. 65, no. 15, pp. 2590–2603, 2020.

[20] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: a large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2141–2149, 2020.