

UNBALANCED SAMPLE DATA PREPROCESSING METHOD AND DEVICE, AND COMPUTER DEVICE

CROSS-REFERENCE TO RELATED APPLICATIONS

[1] The application claims priority to Chinese Patent Application No. 2018110189130 filed to the China National Intellectual Property Administration on September 3, 2018 and entitled "Unbalanced Sample Data Preprocessing Method and Device, and Computer Device ", the disclosure of which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[2] The application relates to an unbalanced sample data preprocessing method and device, and a computer device and a storage medium.

BACKGROUND

[3] With the development of computer technology and social informationization, it is necessary to process massive data in practical production and application, and most of the original data obtained belong to unbalanced data, that is, in the whole data set sample space, there is a huge difference in the number of samples of one type and the rest or several types.

[4] At present, in the case of sample imbalance, that is, when there is a large difference in the amount of data corresponding to two types of labels, for the dataset with a small sample size, oversampling is usually used to balance the sample. The traditional oversampling algorithm, for example, the SMOTE algorithm, gives new data between (0, 1) according to $X_new = X + \text{rand}(0, 1) * (X_mean - X)$.

[5] However, the new data generated by the current traditional processing method with sample imbalance cannot meet the actual data standards, and there is still a problem of sample imbalance.

SUMMARY

[6] According to various embodiments disclosed in the application, an unbalanced sample data preprocessing method and device, and a computer device and a storage medium are provided.

[7] An unbalanced sample data preprocessing method includes the following.

[8] A data acquisition request is received, and initial data is acquired according to the

data acquisition request.

[9] The initial data is classified according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold.

[10] K first sample points in the first-class sample sets are extracted.

[11] Characteristics of the K first sample points are analyzed to obtain a new data characteristic of the first-class sample sets.

[12] A first-class label corresponding to the first-class sample sets is obtained, and a new data label of the first-class sample sets is generated according to the first-class label.

[13] The number of first-class sample sets and the number of second-class sample sets are respectively obtained, and a ratio between the number of first-class sample sets and the number of second-class sample sets is calculated.

[14] New data of the first-class sample sets is generated according to the new data characteristic and the new data label, and the amount of new data is adjusted according to the ratio to increase the number of first-class sample sets.

[15] An unbalanced sample data preprocessing device includes: an initial data acquiring module, a classifying module, an extracting module, a new data characteristic obtaining module, a new data label generating module, a ratio calculating module and a new data generating module.

[16] The initial data acquiring module is configured to receive a data acquisition request, and acquire initial data according to the data acquisition request.

[17] The classifying module is configured to classify the initial data according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold.

[18] The extracting module is configured to extract K first sample points in the first-class sample sets.

[19] The new data characteristic obtaining module is configured to analyze characteristics of the K first sample points to obtain a new data characteristic of the first-class sample sets.

[20] The new data label generating module is configured to obtain a first-class label corresponding to the first-class sample sets, and generate a new data label of the first-class

sample sets according to the first-class label.

[21] The ratio calculating module is configured to respectively obtain the number of first-class sample sets and the number of second-class sample sets, and calculate a ratio between the number of first-class sample sets and the number of second-class sample sets.

[22] The new data generating module is configured to generate new data of the first-class sample sets according to the new data characteristic and the new data label, and adjust the amount of new data according to the ratio to increase the number of first-class sample sets.

[23] A computer device includes a memory and one or more processors. The memory stores a computer readable instruction. The computer readable instruction, when executed by one or more processors, causes the one or more processors to perform the following steps.

[24] A data acquisition request is received, and initial data is acquired according to the data acquisition request.

[25] The initial data is classified according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold.

[26] K first sample points in the first-class sample sets are extracted.

[27] Characteristics of the K first sample points are analyzed to obtain a new data characteristic of the first-class sample sets.

[28] A first-class label corresponding to the first-class sample sets is obtained, and a new data label of the first-class sample sets is generated according to the first-class label.

[29] The number of first-class sample sets and the number of second-class sample sets are respectively obtained, and a ratio between the number of first-class sample sets and the number of second-class sample sets is calculated.

[30] New data of the first-class sample sets is generated according to the new data characteristic and the new data label, and the amount of new data is adjusted according to the ratio to increase the number of first-class sample sets.

[31] One or more non-transitory computer readable storage media storing a computer readable instruction are provided. The computer readable instruction, when executed by one or more processors, causes the one or more processors to perform the following steps.

[32] A data acquisition request is received, and initial data is acquired according to the data acquisition request.

[33] The initial data is classified according to a preset classification rule to obtain

first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold.

[34] K first sample points in the first-class sample sets are extracted.

[35] Characteristics of the K first sample points are analyzed to obtain a new data characteristic of the first-class sample sets.

[36] A first-class label corresponding to the first-class sample sets is obtained, and a new data label of the first-class sample sets is generated according to the first-class label.

[37] The number of first-class sample sets and the number of second-class sample sets are respectively obtained, and a ratio between the number of first-class sample sets and the number of second-class sample sets is calculated.

[38] New data of the first-class sample sets is generated according to the new data characteristic and the new data label, and the amount of new data is adjusted according to the ratio to increase the number of first-class sample sets.

[39] Details of one or more embodiments of the application are presented in the drawings and descriptions below. Other features and advantages of the application will become apparent in the description, the drawings and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[40] In order to more clearly illustrate technical solutions in embodiments of the application, the accompanying drawings needed in the embodiments are simply introduced below. It is apparent that the accompanying drawings in the following description are only some embodiments of the application, for the ordinary skill in the art, some other accompanying drawings can also be obtained according to these on the premise of not contributing creative effort.

[41] Fig. 1 is a schematic diagram of an application scenario of an unbalanced sample data preprocessing method according to one or more embodiments.

[42] Fig. 2 is a flowchart of an unbalanced sample data preprocessing method according to one or more embodiments.

[43] Fig. 3 is a flowchart of analyzing characteristics of K first sample points to obtain a new data characteristic according to one or more embodiments.

[44] Fig. 4 is a block diagram of an unbalanced sample data preprocessing device according to one or more embodiments.

[45] Fig. 5 is a block diagram of a computer device according to one or more

embodiments.

DETAILED DESCRIPTION

[46] In order to make the technical solutions and advantages of the application clearer, the application will further be described below in combination with the drawings and embodiments in detail. It should be understood that the specific embodiments described here are only used to explain the application and are not intended to limit the application.

[47] An unbalanced sample data preprocessing method provided by the application may be applied in the application environment as shown in Fig. 1. A terminal 102 communicates with a server 104 through a network. The server 104 receives a data acquisition request sent by the terminal 102, and acquires initial data according to the data acquisition request. The initial data is classified according to a preset classification rule to obtain first-class sample sets and second-class sample sets. The number of samples in each of the first-class sample sets is less than a data amount threshold, and the number of samples in each of the second-class sample sets is greater than the data amount threshold. K first sample points in the first-class sample sets are extracted, and characteristics of K first sample points are analyzed to obtain a new data characteristic of the first-class sample sets. A first-class label corresponding to the first-class sample sets is obtained, and a new data label of the first-class sample sets is generated according to the first-class label. The number of first-class sample sets and the number of second-class sample sets are respectively obtained, and a ratio between the number of first-class sample sets and the number of second-class sample sets is calculated. New data of the first-class sample sets is generated according to the new data characteristic and the new data label, and the amount of new data is adjusted according to the ratio to increase the number of first-class sample sets. The terminal 102 may be, but not limited to, a personal computer, a laptop, a smart phone, a tablet computer and a portable wearable device. The server 104 may be realized by an independent server or a cluster of multiple servers.

[48] In one of the embodiments, as shown in Fig. 2, an unbalanced sample data preprocessing method is provided. Illustrated by the application of the method to the server in Fig. 1, the method includes the following steps.

[49] At S202, a server receives a data acquisition request sent by a terminal, and acquires initial data according to the data acquisition request.

[50] Specifically, different data acquisition requests correspond to different initial data, and the server may acquire the initial data corresponding to the data acquisition request sent by the terminal according to a preset corresponding relationship between the data

acquisition request and the initial data. The data acquisition request carries a data identity. The server may acquire the initial data corresponding to the data identity according to the corresponding relationship between the data identity and the initial data.

[51] Taking a modeling stage for example, the terminal is required to acquire a mass of sample data to build a model. The server receives a modeling data acquisition request sent by the terminal and extracts a modeling data identity carried in the modeling data acquisition request, obtains a preset corresponding relationship between the modeling data identity and the initial data, and acquires the initial data corresponding to the modeling data identity.

[52] The server acquires original data from a database according to the data acquisition request, and obtains data types corresponding to the original data. A preset corresponding relationship between the data types and processing modes is obtained, and the original data corresponding to each data type is processed to obtain the original data.

[53] At S204, the server classifies the initial data according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold.

[54] Specifically, the preset classification rule is a rule that the initial data is classified according to a data amount threshold. The server classifies the initial data with fewer samples than the data amount threshold into the first-class sample sets, and classifies the initial data with more samples than the data amount threshold into the second-class sample sets.

[55] Taking a modeling stage for example, most of the available initial data belong to unbalanced data, that is, in the whole data set sample space, there is a huge difference in the number of samples of one type and the rest or several types. In the embodiment, the server may classify the obtained initial data according to a preset data amount threshold to obtain the first-class sample sets, that is, the initial data with fewer samples than the data amount threshold. The server may also classify the initial data with more samples than the data amount threshold to obtain the second-class sample sets.

[56] The first-class sample set includes multiple first-class samples, there are multiple first sample points in the first-class sample, and the number of first-class samples is less than the preset data amount threshold, so the first-class sample set may also be called few-class sample set, its corresponding sample may also be called few-class sample, and the first sample point may also be called few-class sample point. The second-class sample set includes multiple second-class samples, there are multiple second sample points in the second-class sample, and the number of second-class samples is greater than the preset data amount

threshold, so the second-class sample set may also be called many-class sample set, its corresponding sample may also be called many-class sample, and the second sample point may also be called many-class sample point.

[57] At S206, the server extracts K first sample points in the first-class sample sets.

[58] Specifically, the first-class sample set includes multiple first-class samples, and the first-class sample includes multiple first sample points. The server extracts a sample set from the first-class sample sets randomly to obtain a first sample set. The distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets are calculated respectively, and the first sample points are sorted according to the distances to obtain the first K first sample points.

[59] The server respectively calculates the Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets, sorts the Euclidean distances in order from small to large, and obtains the first sample points corresponding to the first K Euclidean distances according to the sorting of the Euclidean distances.

[60] The Euclidean distance is the true distance between two points in an m-dimensional space, or the natural length of a vector, that is, the distance from that point to the origin. The Euclidean distance in a two-dimensional space and a three-dimensional space is the actual distance between two points. The Euclidean distances calculated are sorted in order from small to large, and the first sample points corresponding to the first K Euclidean distances are obtained, that is, the obtained first sample points are the sample points corresponding to the first K minimum Euclidean distances.

[61] At S208, the server analyzes characteristics of K first sample points to obtain a new data characteristic of the first-class sample sets.

[62] Specifically, the server extracts the characteristics of K first sample points and analyzes the characteristics to obtain a characteristic attribute. The characteristics of the K first sample points are respectively extracted according to the characteristic attribute, and common characteristics of the K first sample points are respectively obtained. Corresponding common characteristic combinations are formed according to the common characteristics, the number of common characteristics included in the common characteristic combinations is calculated, and the common characteristic combinations are sorted according to the number of common characteristics to obtain a common characteristic combination corresponding to the maximum number. The new data characteristic is generated according to the common characteristic combination corresponding to the maximum number. The characteristic

attribute refers to the attributes of characteristic about the first sample point, including the category of the characteristic and the definition of the characteristic, obtained after the characteristic of the first sample point is analyzed and processed. Characteristic extraction refers to using a computer to extract image information to determine whether points each image belong to an image characteristic.

[63] The server forms the corresponding common characteristic combination according to the obtained common characteristics. The number of common characteristics included in the different common characteristic combinations is calculated, and the common characteristic combinations are sorted according to the calculated number of common characteristics to obtain a common characteristic combination corresponding to the maximum number. For example, if the number of common characteristics included in the different common characteristic combinations is 4, 5, 3, 6 and 2, and the number of common characteristics obtained after sorting is 6, 5, 4, 3 and 2, then the obtained common characteristic combination corresponding to the maximum number is the common characteristic combination corresponding to the number of common characteristics 6. The server obtains the calculated common characteristic combination corresponding to the maximum number, obtains the common characteristics included in the common characteristic combination, and takes the multiple common characteristics as the new data characteristic. The new data characteristic is the characteristic corresponding to the sample data included in the new first-class sample set.

[64] At S210, the server obtains a first-class label corresponding to the first-class sample sets, and generates a new data label of the first-class sample sets according to the first-class label.

[65] Specifically, the server obtains the first-class sample in the first-class sample set, and obtains the first-class label corresponding to the first-class sample according to a preset corresponding relationship between the first-class sample and the first-class label. The first-class sample set includes multiple first-class samples. Different first-class samples correspond to different first-class labels. The server may respectively obtain the first-class labels corresponding to different first-class samples according to the corresponding relationship between the first-class sample and the first-class label, and generate the new data label according to the obtained first-class label. The new data label represents a label carried in the data newly added for the first-class sample set. The new data of the first-class sample set may be generated according to the new data label and the new data characteristic.

[66] At S212, the server obtains the number of first-class sample sets and the number of

second-class sample sets respectively, and calculates a ratio between the number of first-class sample sets and the number of second-class sample sets.

[67] At S214, the server generates new data of the first-class sample sets according to the new data characteristic and the new data label, and adjusts the amount of new data according to the ratio to increase the number of first-class sample sets.

[68] Specifically, the server obtains the number of first sample sets, that is, the number of sample points in the first sample set, obtains the number of second sample sets, that is, the number of sample points in the second sample set, calculates the ratio between the number of sample points in the first sample set and the number of sample points in the second sample set, and adjusts the amount of new data according to the ratio calculated, so as to increase the number of sample points in the first-class sample set and achieve a balanced state between the first-class sample set and the second-class sample set. The server generates the new data of the first-class sample set according to the new data characteristic and the new data label, that is, the new data is generated through the new data characteristic and the new data label.

[69] When the number of sample points in the first sample set obtained by the server is 20, and the number of sample points in the second sample set obtained is 50, the ratio between the number of first-class sample sets and the number of second-class sample sets is 2:5 through calculation, and according to the obtained ratio, the amount of the new data is adjusted to generate 30 units of new data to achieve a balance between the first sample set and the second sample set. At the same time, the new data is generated through the new data characteristic and the new data label.

[70] In the unbalanced sample data preprocessing method, the server classifies the initial data according to the preset classification rule to obtain the first-class sample set and the second-class sample set, extracts K first sample points and analyzes the characteristics thereof, generates the new data label of the first-class sample set according to the new data characteristic of the first-class sample set, calculates the ratio between the number of first-class sample sets and the number of second-class sample sets, generates the new data of the first-class sample set through the new data characteristic and the new data label, and adjusts the amount of the new data according to the ratio to increase the number of first-class sample sets, which may make the number of first-class sample sets and the number of second-class sample sets in a balanced state and solve the problem of sample imbalance.

[71] In one of the embodiments, as shown in Fig. 3, the following steps are provided to analyze the characteristics of K first sample points and obtain the new data characteristic.

[72] At S302, the server extracts the characteristics of the K first sample points.

[73] At S304, the server analyzes the characteristics to obtain a characteristic attribute.

[74] Specifically, after extracting the characteristics of K first sample points to obtain the characteristics of the K first sample points, the server analyzes the characteristics to obtain the attributes of characteristic. The attributes of characteristic include the category of the characteristic and the definition of the characteristic. Taking any one of the K first sample points for example, the server extracts the characteristic of this first sample point, analyzes the characteristic and obtains the corresponding attributes of characteristic, including the categories to which the different characteristics belong and the respective definitions of the different characteristics.

[75] At S306, the server extracts the characteristics of the K first sample points respectively according to the characteristic attributes, and obtains common characteristics of the K first sample points respectively.

[76] Specifically, characteristic extraction refers to using a computer to extract image information to determine whether points of each image belong to an image characteristic. The server may obtain the common characteristic of each of the K first sample points through characteristic extraction. By determining the category of the characteristic, the different first sample points are classified according to the category of the characteristic. The server extracts the definition of different characteristics and obtains an intersection of the different characteristics according to the definition of characteristic, and extracts the intersection to obtain the common characteristics among the different characteristics.

[77] At S308, the server forms corresponding common characteristic combinations according to the common characteristics, and calculates the number of common characteristics included in the common characteristic combinations.

[78] At S310, the server sorts the common characteristic combinations according to the number of common characteristics to obtain a common characteristic combination corresponding to the maximum number.

[79] Specifically, the server forms the corresponding common characteristic combination according to the obtained common characteristics. The number of common characteristics included in the different common characteristic combinations is calculated, and the common characteristic combinations are sorted according to the calculated number of common characteristics to obtain the common characteristic combination corresponding to the maximum number. For example, if the number of common characteristics included in the different common characteristic combinations is 2, 7, 4, 6 and 9, and the number of common characteristics obtained after sorting is 9, 7, 6, 4 and 2, then the obtained common

characteristic combination corresponding to the maximum number is the common characteristic combination corresponding to the number of common characteristics 9.

[80] At S312, the server generates the new data characteristic according to the common characteristic combination corresponding to the maximum number.

[81] Specifically, the server obtains the calculated common characteristic combination corresponding to the maximum number, obtains the common characteristics included in the common characteristic combination, and takes the multiple common characteristics as the new data characteristic. The new data characteristic is the characteristic corresponding to the sample data included in the new first-class sample set.

[82] In the step of analyzing the characteristics of K first sample points to obtain the new data characteristic, the server analyzes the characteristics of the K first sample points to obtain the characteristic attributes, and extracts the characteristics of the K first sample points respectively according to the characteristic attributes to obtain the common characteristics of the K first sample points respectively. Corresponding common characteristic combinations are formed according to the common characteristics, the number of common characteristics included in the common characteristic combinations is calculated, and the common characteristic combinations are sorted according to the number of common characteristics to obtain a common characteristic combination corresponding to the maximum number. By generating the new data characteristic according to the common characteristic combination corresponding to the maximum number, the generated new data characteristic may, to the greatest extent, include the characteristic corresponding to the first-class sample set, so as to ensure the similarity between the new data and the sample points in the first set.

[83] In one of the embodiments, a step of receiving a data acquisition request and acquiring initial data according to the data acquisition request is provided, which includes:

[84] the server acquires original data from a database according to a data acquisition request, obtains data type corresponding to the original data, obtains a preset corresponding relationship between the data type and a processing mode, and according to the processing mode, processes the original data corresponding to each data type to obtain the initial data.

[85] Specifically, different original data corresponds to different data types. The data types include digital type, byte type and text type, and the corresponding original data includes digital original data, byte original data and text original data. The server may obtain the data type corresponding to the original data according to the preset corresponding relationship between the data type and the original data. Different data types correspond to different processing modes. The data types include digital type, byte type and text type, and

the corresponding processing modes are judgment processing, assignment processing, and declaration processing.

[86] Further, for the digital original data, judgment processing is performed, that is, a preset value range is obtained, the preset value range is compared with the value of the digital original data, it is determined whether the value of the digital original data conforms to the preset value range, and the digital original data conforming to the preset value range is extracted to generate the digital original data. For the byte original data, assignment processing is performed, that is, it is determined whether the value of the byte original data conforms to a preset value, and when the value of the byte original data does not conform to the preset value, the preset value is assigned to the corresponding byte original data, and byte initial data is generated according to the byte original data after assignment. For the text original data, declaration processing is performed, that is, a component of the text original data is obtained and compared with a preset component, and when the component of the text original data is inconsistent with the preset component, the text original data is declared as the preset component.

[87] In the step of receiving the data acquisition request and acquiring the initial data according to the data acquisition request, by processing the original data according to the data type and generating the initial data targeted to different data types, the work efficiency is improved.

[88] In one of the embodiments, a step of classifying the initial data according to the preset classification rule to obtain the first-class sample set and the second-class sample set is provided, which includes:

[89] the server obtains the sample set corresponding to each data type according to the data type corresponding to the initial data, calculates the amount of the initial data of the sample set, compares the data amount of each sample set with the preset data amount threshold, and when the data amount exceeds the data amount threshold, obtains the second-class sample set according to the corresponding sample set, and when the data amount is lower than the data amount threshold, obtains the first-class sample set according to the corresponding sample set.

[90] Specifically, the server obtains the data type corresponding to the initial data according to the preset corresponding relationship between the initial data and the data type, and obtains the sample set corresponding to each data type. The server calculates the data amount corresponding to the initial data of the sample set, and compares the data amount of each sample set with the preset data amount threshold to determine whether the data amount

of each sample set exceeds the preset data amount threshold. When the data amount exceeds the data amount threshold, the sample set that the data amount exceeds the data amount threshold is the second-class sample set. When the data amount does not exceed the data amount threshold, the sample set that the data amount does not exceed the data amount threshold is the first-class sample set.

[91] The data types include digital type, byte type and text type, and the corresponding sample sets are digital sample set, byte sample set and text sample set. The server may obtain the corresponding digital data type according to the digital initial data, and obtain the corresponding digital sample set. Similarly, the server may obtain the corresponding byte data type and the text data type respectively according to the byte initial data and the text initial data, and then obtain the corresponding byte sample set and text sample set.

[92] Further, the number of data amount thresholds may be set to 100, and the data amounts of each sample set obtained are 80, 95, 108 and 120, so the sample sets with 80 and 95 data amounts are the first-class sample sets, while the sample sets with 108 and 120 data amounts are the second-class sample sets.

[93] In the step of classifying the initial data according to the preset classification rule to obtain the first-class sample set and the second-class sample set, by classifying the sample sets in advance according to the data amount to obtain the first-class sample set and the second-class sample set, the sample set that needs the new data may be learned as soon as possible, the process of data addition is speeded up, and the work efficiency is improved.

[94] In one of the embodiments, a step of extracting K first sample points in the first-class sample set is provided, which includes:

[95] the server extracts a sample set from the first-class sample sets randomly to obtain the first sample set, respectively calculates the distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets, and sorts the first sample points according to the distances to obtain the first K first sample points.

[96] Specifically, the first-class sample set includes multiple first-class samples, and the first-class sample includes multiple first sample points. The server extracts a sample set from the first-class sample sets randomly to obtain a first sample set. The distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets are calculated respectively, and the first sample points are sorted in the order of distances from small to large to obtain the first K first sample points.

[97] In the step of extracting K first sample points in the first-class sample set, by

calculating the distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets, and obtaining the first sample points corresponding to first K distances, the extracted first sample point is ensured to be the sample point with the smallest distance in the first K distances, and the accuracy of sample point extraction is improved.

[98] In one of the embodiments, a step of extracting K first sample points in the first-class sample set is provided, which further includes:

[99] the server respectively calculates the Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets, sorts the Euclidean distances in order from small to large, and obtains the first sample points corresponding to the first K Euclidean distances according to the sorting of the Euclidean distances.

[100] Specifically, the Euclidean distance represents the true distance between two points in a m-dimensional space, or the natural length of a vector, that is, the distance from that point to the origin. The server sorts the Euclidean distances calculated in order from small to large, and obtains the first sample points corresponding to the first K Euclidean distances, that is, the obtained first sample points are the sample points corresponding to the first K minimum Euclidean distances.

[101] In the step of extracting K first sample points in the first-class sample set, the server respectively calculates the Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets, and sorts the Euclidean distances in order from small to large. According to the sorting of Euclidean distances, the first sample points corresponding to the first K Euclidean distances are obtained, which further ensures that the extracted first sample points are the sample points with the lowest one of the first K Euclidean distances, and the accuracy of sample point extraction is improved.

[102] It should be understood that although the steps in the flowcharts in Fig. 2 to Fig. 3 are shown in order as indicated by the arrows, they are not necessarily performed in the order indicated by the arrows. Unless explicitly stated in the application, there is no strict order in which these steps are performed, and they can be performed in any other order. Furthermore, at least a part of steps in Fig. 2 to Fig. 3 may include multiple sub-steps or multiple stages. These sub-steps or phases are not necessarily performed at the same time, but may be performed at different times. These sub-steps or stages are not necessarily performed in order, but may be performed in turn or alternately with at least a part of other steps or sub-steps or phases of the other steps.

[103] In one of the embodiments, as shown in Fig. 4, an unbalanced sample data preprocessing device is provided, which includes: an initial data acquiring module 402, a classifying module 404, an extracting module 406, a new data characteristic obtaining module 408, a new data label generating module 410, a ratio calculating module 412 and a new data generating module 414.

[104] The initial data acquiring module 402 is configured to receive the data acquisition request, and acquire the initial data according to the data acquisition request.

[105] The classifying module 404 is configured to classify the initial data according to the preset classification rule to obtain the first-class sample set and the second-class sample set. The number of samples in each of the first-class sample sets is less than the data amount threshold, and the number of samples in each of the second-class sample sets is greater than the data amount threshold.

[106] The extracting module 406 is configured to extract K first sample points in the first-class sample sets.

[107] The new data characteristic obtaining module 408 is configured to analyze the characteristics of the K first sample points to obtain the new data characteristic of the first-class sample sets.

[108] The new data label generating module 410 is configured to obtain the first-class label corresponding to the first-class sample sets, and generate the new data label of the first-class sample sets according to the first-class label.

[109] The ratio calculating module 412 is configured to respectively obtain the number of first-class sample sets and the number of second-class sample sets, and calculate a ratio between the number of first-class sample sets and the number of second-class sample sets.

[110] The new data generating module 414 is configured to generate the new data of the first-class sample sets according to the new data characteristic and the new data label, and adjust the amount of new data according to the ratio to increase the number of first-class sample sets.

[111] With the unbalanced sample data preprocessing device, the server classifies the initial data according to the preset classification rule to obtain the first-class sample set and the second-class sample set, extracts K first sample points and analyzes the characteristics thereof, generates the new data label of the first-class sample set according to the new data characteristic of the first-class sample set, calculates the ratio between the number of first-class sample sets and the number of second-class sample sets, generates the new data of the first-class sample set through the new data characteristic and the new data label, and

adjusts the amount of the new data according to the ratio to increase the number of first-class sample sets, which may make the number of first-class sample sets and the number of second-class sample sets in a balanced state and solve the problem of sample imbalance.

[112] In one of the embodiments, an initial data acquiring module is provided, which is further configured to: acquire the original data from the database according to the data acquisition request; obtain the data type corresponding to the original data; obtain the preset corresponding relationship between the data type and the processing mode; and according to the processing mode, process the original data corresponding to each data type to obtain the initial data.

[113] With the initial data acquiring module, by processing the original data according to the data type and generating the initial data targeted to different data types, the work efficiency is improved.

[114] In one of the embodiments, a classifying module is provided, which is further configured to: obtain the sample set corresponding to each data type according to the data type corresponding to the initial data; calculate the amount of the initial data of the sample set; compare the data amount of each sample set with the preset data amount threshold; when the data amount exceeds the data amount threshold, obtain the second-class sample set according to the corresponding sample set; and when the data amount is lower than the data amount threshold, obtain the first-class sample set according to the corresponding sample set.

[115] With the classifying module, by classifying the sample sets in advance according to the data amount to obtain the first-class sample set and the second-class sample set, the sample set that needs the new data may be learned as soon as possible, the process of data addition is speeded up, and the work efficiency is improved.

[116] In one of the embodiments, an extracting module is provided, which is further configured to: extract a sample set from the first-class sample sets randomly to obtain the first sample set; respectively calculate the distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets; and sort the first sample points according to the distances to obtain the first K first sample points.

[117] With the extracting module, by calculating the distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets, and obtaining the first sample points corresponding to first K distances, the extracted first sample point is ensured to be the sample point with the smallest distance in the first K distances, and the accuracy of sample point extraction is improved.

[118] In one of the embodiments, an extracting module is provided, which is further

configured to: respectively calculate the Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets; sort the Euclidean distances in order from small to large; and obtain the first sample points corresponding to the first K Euclidean distances according to the ordering of the Euclidean distances.

[119] With the extracting module, the server respectively calculates the Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets, and sorts the Euclidean distances in order from small to large. According to the sorting of Euclidean distances, the first sample points corresponding to the first K Euclidean distances are obtained, which further ensures that the extracted first sample points are the sample points with the lowest one of the first K Euclidean distances, and the accuracy of sample point extraction is improved.

[120] In one of the embodiments, a new data generating module is provided, which is further configured to: extract the characteristics of K first sample points; analyze the characteristics to obtain the characteristic attributes; respectively extract the characteristics of the K first sample points according to the characteristic attributes to respectively obtain the common characteristics of the K first sample points; form the common characteristic combinations according to the common characteristics and calculate the number of common characteristics included in the common characteristic combinations; sort the common characteristic combinations according to the number of common characteristics to obtain a common characteristic combination corresponding to the maximum number; and generate the new data characteristic according to the common characteristic combination corresponding to the maximum number.

[121] With the new data generating module, the server analyzes the characteristics of the K first sample points to obtain the characteristic attributes, and extracts the characteristics of the K first sample points respectively according to the characteristic attributes to obtain the common characteristics of the K first sample points respectively. The corresponding common characteristic combinations are formed according to the common characteristics, the number of common characteristics included in the common characteristic combinations is calculated, and the common characteristic combinations are sorted according to the number of common characteristics to obtain a common characteristic combination corresponding to the maximum number. By generating the new data characteristic according to the common characteristic combination corresponding to the maximum number, the generated new data characteristic may, to the greatest extent, include the characteristic corresponding to the first-class sample set, so as to ensure the similarity between the new data and the sample points in the first set.

[122] For the specific limitations of the unbalanced sample data preprocessing device, the above limitations of the unbalanced sample data preprocessing method may be referred to, which will not be repeated here. Each module in the unbalanced sample data preprocessing device may be realized in whole or in part by software, hardware and their combination. Each above module may be embedded in or independent of a processor in a computer device in the form of hardware, or stored in a memory in the computer device in the form of software, so that the processor may call and perform the operation corresponding to each module above.

[123] In one of the embodiments, a computer device is provided. The computer device may be a server, and its internal structure may be shown in Fig. 5. The computer device includes a processor, a memory, a network interface and a database connected through a system bus. The processor of the computer device is used to provide computing and control capabilities. The memory of the computer device includes a non-transitory computer readable storage medium and an internal memory. The non-transitory computer readable storage medium stores an operating system, a computer readable instruction and a database. The internal memory provides an environment for the operation of the operating system and the computer readable instruction in the non-transitory computer readable storage medium. The database of the computer device is used to store unbalanced sample data. The network interface of the computer device is used to communicate with an external terminal through a network connection. The computer readable instruction, when executed by the processor, performs an unbalanced sample data preprocessing method.

[124] Those skilled in the art may understand that the structure shown in Fig. 5 is only a block diagram of part of the structure related to the solutions of the application and does not constitute a limitation on the computer device applied to the solutions of the application. Specifically, the computer device may include more or fewer parts than that are shown in the figures, or some combination of parts, or a different arrangement of parts. A computer device includes a memory and one or more processors. The memory stores a computer readable instruction. The computer readable instruction, when executed by the processor, performs the steps of the unbalanced sample data preprocessing method provided in any embodiment of the application.

[125] One or more non-transitory computer readable storage media storing a computer readable instruction are provided. The computer readable instruction, when executed by one or more processors, causes the one or more processors to perform the steps of the unbalanced sample data preprocessing method provided in any embodiment of the application.

[126] Those of ordinary skill in the art may understand that all or a part of flows of the

method in the above embodiments may be completed by related hardware instructed by a computer readable instruction. The computer readable instruction may be stored in a non-transitory computer readable storage medium. When executed, the computer readable instruction may include the flows in the embodiments of the method. Any reference to memory, storage, database or other media used in each embodiment provided in the application may include non-transitory and/or transitory memories. The non-transitory memories may include a Read-Only Memory (ROM), a Programmable Read-Only Memory (PROM), an Electrically Programmable Read-Only Memory (EPROM), an Electrically Erasable Programmable Read-Only Memory (EEPROM) or a flash memory. The transitory memories may include a Random Access Memory (RAM) or an external cache memory. As an illustration rather than a limitation, the RAM is available in many forms, such as Static RAM (SRAM), Dynamic RAM (DRAM), Synchronous DRAM (SDRAM), Double Data Rate SDRAM (DDRAM), Enhanced SDRAM (ESDRAM), Synch-link DRAM (SLDRAM), Rambus Direct RAM (RDRAM), Direct Memory Bus Dynamic RAM (DRDRAM), and Memory Bus Dynamic RAM (RDRAM).

[127] The technical characteristics of the above embodiments may be combined at random. To make the description concise, not all possible combinations of these technical characteristics of the above embodiments are described, however, all the combinations of these technical characteristics shall fall within the scope of the description, as long as there is no contradiction in the combinations of these technical characteristics.

[128] The above embodiments only express several implementation modes of the application. The descriptions of these embodiments are specific and detailed, however, they cannot be understood as a limitation to the claimed scope of the disclosure. It should be pointed out that those of ordinary skill in the art may also make several improvements and modifications without departing from the conception of the application, and these improvements and modifications should fall within the protection scope of protection of the application. Therefore, the protection scope of the application is subject to the attached claims.

What is claimed is:

1. An unbalanced sample data preprocessing method, comprising:

receiving a data acquisition request, and acquiring initial data according to the data acquisition request;

classifying the initial data according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold;

extracting K first sample points in the first-class sample sets;

analyzing characteristics of the K first sample points to obtain a new data characteristic of the first-class sample sets;

obtaining a first-class label corresponding to the first-class sample sets, and generating a new data label of the first-class sample sets according to the first-class label;

respectively obtaining the number of first-class sample sets and the number of second-class sample sets, and calculating a ratio between the number of first-class sample sets and the number of second-class sample sets; and

generating new data of the first-class sample sets according to the new data characteristic and the new data label, and adjusting the amount of new data according to the ratio to increase the number of first-class sample sets.

2. The method as claimed in claim 1, wherein receiving the data acquisition request, and acquiring the initial data according to the data acquisition request comprises:

acquiring original data from a database according to the data acquisition request;

obtaining data types corresponding to the original data;

obtaining a preset corresponding relationship between a data type and a processing mode; and

processing original data corresponding to each data type according to the processing mode to obtain the initial data.

3. The method as claimed in claim 1, wherein classifying the initial data according to the preset classification rule to obtain multiple first-class sample sets and second-class sample sets comprises:

obtaining a sample set corresponding to each data type according to the data types corresponding to the initial data;

calculating data amount of the initial data of the sample set;

comparing the data amount of each sample set with a preset data amount threshold; when

the data amount exceeds the data amount threshold, obtaining the second-class sample sets according to the corresponding sample set; and

when the data amount is lower than the data amount threshold, obtaining the first-class sample sets according to the corresponding sample set.

4. The method as claimed in any one of claims 1 to 3, wherein extracting the K first sample points in the first-class sample sets comprises:

extracting a sample set from the first-class sample sets randomly to obtain a first sample set;

respectively calculating distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets; and

sorting first sample points according to the distances to obtain the first K first sample points.

5. The method as claimed in claim 4, further comprising:

respectively calculating Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets; and

sorting the Euclidean distances in order from small to large; and

obtaining first sample points corresponding to the first K Euclidean distances according to the sorting of the Euclidean distances.

6. The method as claimed in any one of claims 1 to 3, wherein analyzing the characteristics of the K first sample points to obtain the new data characteristic comprises:

extracting the characteristics of the K first sample points;

analyzing the characteristics to obtain a characteristic attribute;

respectively extracting the characteristics of the K first sample points according to the characteristic attribute, and respectively obtaining common characteristics of the K first sample points;

forming corresponding common characteristic combinations according to the common characteristics, and calculating the number of common characteristics comprised in the common characteristic combinations;

sorting the common characteristic combinations according to the number of common characteristics to obtain a common characteristic combination corresponding to the maximum number; and

generating the new data characteristic according to the common characteristic combination corresponding to the maximum number.

7. An unbalanced sample data preprocessing device, comprising:

an initial data acquiring module, configured to receive a data acquisition request, and acquire initial data according to the data acquisition request;

a classifying module, configured to classify the initial data according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold;

an extracting module, configured to extract K first sample points in the first-class sample sets;

a new data characteristic obtaining module, configured to analyze characteristics of the K first sample points to obtain a new data characteristic of the first-class sample sets;

a new data label generating module, configured to obtain a first-class label corresponding to the first-class sample sets, and generate a new data label of the first-class sample sets according to the first-class label;

a ratio calculating module, configured to respectively obtain the number of first-class sample sets and the number of second-class sample sets, and calculate a ratio between the number of first-class sample sets and the number of second-class sample sets; and

a new data generating module, configured to generate new data of the first-class sample sets according to the new data characteristic and the new data label, and adjust the amount of new data according to the ratio to increase the number of first-class sample sets.

8. The device as claimed in claim 7, wherein the initial data acquiring module is further configured to:

acquire original data from a database according to the data acquisition request;

obtain data types corresponding to the original data;

obtain a preset corresponding relationship between a data type and a processing mode;

and

process original data corresponding to each data type according to the processing mode to obtain the initial data.

9. The device as claimed in claim 7, wherein the classifying module is further configured to:

obtain a sample set corresponding to each data type according to the data types corresponding to the initial data;

calculate data amount of the initial data of the sample set;

compare the data amount of each sample set with a preset data amount threshold; when

the data amount exceeds the data amount threshold, obtain the second-class sample sets according to the corresponding sample set; and

when the data amount is lower than the data amount threshold, obtain the first-class sample sets according to the corresponding sample set.

10. A computer device, comprising: a memory and one or more processors, the memory storing a computer readable instruction; the computer readable instruction, when executed by one or more processors, causes the one or more processors to perform the following steps:

receiving a data acquisition request, and acquiring initial data according to the data acquisition request;

classifying the initial data according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold;

extracting K first sample points in the first-class sample sets;

analyzing characteristics of the K first sample points to obtain a new data characteristic of the first-class sample sets;

obtaining a first-class label corresponding to the first-class sample sets, and generating a new data label of the first-class sample sets according to the first-class label;

respectively obtaining the number of first-class sample sets and the number of second-class sample sets, and calculating a ratio between the number of first-class sample sets and the number of second-class sample sets; and

generating new data of the first-class sample sets according to the new data characteristic and the new data label, and adjusting the amount of new data according to the ratio to increase the number of first-class sample sets.

11. The computer device as claimed in claim 10, wherein the processor, when executing the computer readable instruction, is further configured to perform:

acquiring original data from a database according to the data acquisition request;

obtaining data types corresponding to the original data;

obtaining a preset corresponding relationship between a data type and a processing mode;

and

processing original data corresponding to each data type according to the processing mode to obtain the initial data.

12. The computer device as claimed in claim 10, wherein the processor, when executing the computer readable instruction, is further configured to perform:

obtaining a sample set corresponding to each data type according to the data types corresponding to the initial data;

calculating data amount of the initial data of the sample set;

comparing the data amount of each sample set with a preset data amount threshold; when the data amount exceeds the data amount threshold, obtaining the second-class sample sets according to the corresponding sample set; and

when the data amount is lower than the data amount threshold, obtaining the first-class sample sets according to the corresponding sample set.

13. The computer device as claimed in any one of claims 10 to 12, wherein the processor, when executing the computer readable instruction, is further configured to perform:

extracting a sample set from the first-class sample sets randomly to obtain a first sample set;

respectively calculating distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets; and

sorting first sample points according to the distances to obtain the first K first sample points.

14. The computer device as claimed in claim 13, wherein the processor, when executing the computer readable instruction, is further configured to perform:

respectively calculating Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets; and

sorting the Euclidean distances in order from small to large; and

obtaining first sample points corresponding to the first K Euclidean distances according to the sorting of the Euclidean distances.

15. The computer device as claimed in any one of claims 10 to 12, wherein the processor, when executing the computer readable instruction, is further configured to perform:

extracting the characteristics of the K first sample points;

analyzing the characteristics to obtain a characteristic attribute;

respectively extracting the characteristics of the K first sample points according to the characteristic attribute, and respectively obtaining common characteristics of the K first sample points;

forming corresponding common characteristic combinations according to the common characteristics, and calculating the number of common characteristics comprised in the common characteristic combinations;

sorting the common characteristic combinations according to the number of common

characteristics to obtain a common characteristic combination corresponding to the maximum number; and

generating the new data characteristic according to the common characteristic combination corresponding to the maximum number.

16. One or more non-transitory computer readable storage media storing a computer readable instruction, the computer readable instruction, when executed by one or more processors, causes the one or more processors to perform the following steps:

receiving a data acquisition request, and acquiring initial data according to the data acquisition request;

classifying the initial data according to a preset classification rule to obtain first-class sample sets and second-class sample sets, the number of samples in each of the first-class sample sets being less than a data amount threshold, and the number of samples in each of the second-class sample sets being greater than the data amount threshold;

extracting K first sample points in the first-class sample sets;

analyzing characteristics of the K first sample points to obtain a new data characteristic of the first-class sample sets;

obtaining a first-class label corresponding to the first-class sample sets, and generating a new data label of the first-class sample sets according to the first-class label;

respectively obtaining the number of first-class sample sets and the number of second-class sample sets, and calculating a ratio between the number of first-class sample sets and the number of second-class sample sets; and

generating new data of the first-class sample sets according to the new data characteristic and the new data label, and adjusting the amount of new data according to the ratio to increase the number of first-class sample sets.

17. The storage media as claimed in claim 16, wherein the computer readable instruction, when executed by the processor, is further configured to perform:

acquiring original data from a database according to the data acquisition request;

obtaining data types corresponding to the original data;

obtaining a preset corresponding relationship between a data type and a processing mode;

and

processing original data corresponding to each data type according to the processing mode to obtain the initial data.

18. The storage media as claimed in claim 16, wherein the computer readable instruction, when executed by the processor, is further configured to perform:

obtaining a sample set corresponding to each data type according to the data types corresponding to the initial data;

calculating data amount of the initial data of the sample set;

comparing the data amount of each sample set with a preset data amount threshold; when the data amount exceeds the data amount threshold, obtaining the second-class sample sets according to the corresponding sample set; and

when the data amount is lower than the data amount threshold, obtaining the first-class sample sets according to the corresponding sample set.

19. The storage media as claimed in any one of claims 16 to 18, wherein the computer readable instruction, when executed by the processor, is further configured to perform:

extracting a sample set from the first-class sample sets randomly to obtain a first sample set;

respectively calculating distances from all the first sample points in the first sample set to each sample point in the other sample sets of the first first-class sample sets; and

sorting first sample points according to the distances to obtain the first K first sample points.

20. The storage media as claimed in any one of claims 16 to 18, wherein the computer readable instruction, when executed by the processor, is further configured to perform:

respectively calculating Euclidean distances from all the first sample points in the first sample set to each sample point in the first-class sample sets; and

sorting the Euclidean distances in order from small to large; and

obtaining the first sample points corresponding to the first K Euclidean distances according to the sorting of the Euclidean distances.