

INTELLIGENT DATA ANALYSIS METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM

FIELD OF INVENTION

[0001] The present application relates to the technical field of data prediction, and in particular to an intelligent data analysis method and apparatus, a computer device, and a storage medium.

BACKGROUND OF INVENTION

[0002] With the rapid development of the information age, data prediction technology is also constantly developing. At present, when major scientific research institutions make predictions on medical data, the accuracy of model prediction is low due to the hysteresis of some medical data. For example, for infectious diseases with a certain incubation period (such as chickenpox), when the conditions for disease outbreaks (such as temperature, humidity, etc.) are met, they may outbreak in the next period. As a result, the accuracy of the model prediction is reduced, so that citizens cannot prevent diseases in time, and cannot control the severity of disease outbreaks.

SUMMARY OF THE INVENTION

[0003] The embodiments of the present application provide an intelligent data analysis method and apparatus, a computer device, and a storage medium to solve the problem that the model prediction accuracy rate is low when the data prediction is performed on lagging data.

[0004] An intelligent data analysis method, comprising:

[0005] using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword;

[0006] determining at least one hit entry based on the public opinion data; and the hit entry corresponding to a public opinion factor;

[0007] obtaining medical data within a historical unit time and a public opinion index

corresponding to the hit entry; and the public opinion index carrying a time label;

[0008]taking the public opinion factor and the public opinion index with the time label as first portrait data;

[0009]obtaining original sample data based on the first portrait data and the medical data;

[0010]cleaning the original sample data to obtain sample data to be processed;

[0011]performing lag processing on the sample data to be processed to obtain lagged sample data;

[0012]performing feature expansion processing on the lagged sample data to obtain target sample data; and

[0013]using an improved multi-granularity cascaded random forest algorithm for training the target sample data to obtain a target prediction model; wherein the improved multi-granularity cascaded random forest algorithm comprises a pooling layer that is used to retain data characteristics.

[0014]An intelligent data analysis apparatus, comprising:

[0015]a public opinion data acquisition module, for using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword;

[0016]a hit entry determination module, for determining at least one hit entry based on the public opinion data; and the hit entry corresponding to a public opinion factor;

[0017]a public opinion index acquisition module, for obtaining medical data within a historical unit time and a public opinion index corresponding to the hit entry; and the public opinion index carrying a time label;

[0018]a first portrait data acquisition module, for taking the public opinion factor and the public opinion index with the time label as first portrait data;

[0019]an original sample data acquisition module, for obtaining original sample data based on the first portrait data and the medical data;

[0020]acquisition module of the sample data to be processed, for cleaning the original sample data to obtain sample data to be processed;

[0021]a lagged sample data acquisition module, for performing lag processing on the sample data to be processed to obtain lagged sample data;

[0022] a target sample data acquisition module, for performing feature expansion processing on the lagged sample data to obtain target sample data;

[0023] a target prediction model acquisition module, for using an improved multi-granularity cascaded random forest algorithm for training the target sample data to obtain a target prediction model; wherein the improved multi-granularity cascaded random forest algorithm comprises a pooling layer that is used to retain data characteristics.

[0024] A computer device comprises a memory, a processor, and computer readable instructions stored in the memory and run in the processor, which executes the computer readable instructions to implement the steps of the above data intelligent analysis method.

[0025] A readable storage medium, which stores computer readable instructions, and the steps to implement the above data intelligent analysis method when the computer readable instructions are executed by a processor.

[0026] Details of one or more embodiments of the present application are set out in the appended drawings and descriptions below, and other features and advantages of the present application will become apparent from the specification, appended drawings and claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0027] In order to explain the technical solutions of the embodiments of the present application more clearly. The drawings that need to be used in the description of the embodiments of the present application are briefly introduced as follows. Obviously, the drawings in the following description are only some of the embodiments of the present application. For those of ordinary skill in the art, other drawings can be obtained based on these drawings without creative labor.

[0028] FIG. 1 is an application environment schematic diagram of the intelligent data analysis method in an embodiment of the present application;

[0029] FIG. 2 is a flow chart of the data intelligence analysis method in an embodiment of the present application;

[0030] FIG. 3 is a specific flow chart of step S60 in FIG. 2;

[0031] FIG. 4 is a specific flow chart of step S80 in FIG. 2;
[0032] FIG. 5 is a flow chart of the data intelligence analysis method in an embodiment of the present application;
[0033] FIG. 6 is a specific flow chart of step S90 in FIG. 2;
[0034] FIG. 7 is a specific flow chart of step S92 in FIG. 6;
[0035] FIG. 8 is a schematic diagram of the intelligent data analysis apparatus in an embodiment of the present application; and
[0036] FIG. 9 is a schematic diagram of the computer device in an embodiment of the present application.

DESCRIPTION OF THE EMBODIMENTS

[0037] The technical solutions in the embodiments of the present application will be described clearly and completely in conjunction with the drawings in the embodiments of the present application. Obviously, the described embodiments are part of the embodiments of the present application, not all of the embodiments. Based on the embodiments of the present application, all other embodiments obtained by those of ordinary skill in the art without creative work fall within the protection scope of the embodiments of the present application.

[0038] The intelligent data analysis method provided by the embodiments of the present application can be applied to an intelligent data analysis tool which can train different prediction models according to sample data corresponding to different topics (such as chickenpox, flu, etc.), especially for lagged sample data, it can effectively ensure the accuracy of model prediction. The intelligent data analysis method can be applied in the application environment as shown in FIG. 1, wherein computer device communicates with server through the network. The computer device can be, but is not limited to, various personal computers, notebook computers, smart phones, tablet computers, and portable wearable devices. The server can be implemented as an independent server.

[0039] In one embodiment, as shown in FIG. 2, an intelligent data analysis method is provided, and the method is applied to the server in FIG. 1 as an example for description, including the following steps:

[0040] S10: Use a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword.

[0041] Wherein, the preset keywords are preset keywords related to communicable diseases, such as chickenpox, redness, pruritic herpes and blisters, etc. Public opinion data refers to the text data publicly released by different users in a third-party information platform to reflect the occurrence of social events. Specifically, with the rapid development of the information age, users are more inclined to use various information platforms to query required information, such as querying whether they have a disease based on their symptoms, etc. When a communicable disease (such as chickenpox) breaks out, there will inevitably be greater searches or attention. Therefore, in this embodiment, according to the preset keywords, a crawler tool is used to crawl the public opinion data containing the preset keywords in the third-party information platform (such as Baidu, Weibo or WeChat). It should be noted that some of the preset keywords related to communicable diseases in this embodiment can be present with default keywords, and then the synonyms corresponding to the default keyword are taken to obtain more keywords for crawling, so as to obtain more relevant information, and provide sufficient data set for subsequent model training.

[0042] S20: Determine at least one hit entry based on the public opinion data; wherein the hit entry corresponds to a public opinion factor.

[0043] Specifically, with the rapid development of the information age, users are more inclined to use various information platforms to query required information, such as querying whether they have a disease based on their symptoms, etc. When a communicable disease (such as chickenpox) breaks out, there will inevitably be greater searches or attention. Therefore, in this embodiment, the daily public opinion factors of 20 years in different regions are selected as another part of the portrait data. The public opinion factors include, but are not limited to, chickenpox, redness, pruritic herpes and blisters, etc.

[0044] Wherein, the public opinion data includes at least one original entry (such as a Baidu entry). Specifically, the expert judges whether it is related to chickenpox according to the information contained in each crawled original entry, so as to determine at least one entry that is actually related to chickenpox as a hit entry. Then,

according to the determined hit entry, each hit entry corresponds to a public opinion factor. The public opinion factor refers to at least one factor related to a preset keyword contained in the hit entry, such as chickenpox, redness, pruritic herpes, and blisters.

[0045] S30: Obtain medical data within a historical unit time and a public opinion index corresponding to the hit entry. The public opinion index carries a time label.

[0046] Wherein, the medical data refers to the historical number of patients (i.e. label data) within a historical unit time, such as the past 20 years, in sentinel hospitals in different regions provided by the Center for Disease Control. Understandably, the unit time is the time label, and the unit time can be customized by the user and is not limited here. In this embodiment, the unit time may be one day, one week, one month, one quarter, or one year, etc., which will not be listed here.

[0047] In this embodiment, the unit time is one week as an example. Specifically, obtain the public opinion index and the medical data corresponding to the hit entry in unit time, and each public opinion index carries a time label, and the time label refers to the release time of the hit entry.

[0048] S40: Take the public opinion factor and the public opinion index with the time label as first portrait data.

[0049] Wherein, the first portrait data refers to the public opinion factor and the public opinion index carrying the time label as the feature data for model training. Specifically, when it is necessary to predict whether a disease will break out in a time interval in the future, the time interval can be a week, a month, a quarter, or a year, depending on the predicted time interval, the processing of sample data will be different. Take the time interval as one week for example, public opinion factors (such as chickenpox, redness and herpes) can be used as column labels, and the public opinion index of the Nth week can be used as row labels to establish partial portrait data. Wherein, the public opinion index of the Nth week includes, but is not limited to, the average public opinion index of the Nth week (that is, the average public opinion index of 7 days a week), the maximum public opinion index of the Nth week, and the minimum public opinion index of the Nth week.

[0050] It should be noted that the following table is a schematic diagram of portrait data established according to public opinion factors in this embodiment.

Understandably, the schematic diagram is only an example and is not limited here.

Public Opinion Index for Week N	Redness	Chicken pox	Herpes	
Average public opinion index of the first week	x_1	x_2	x_3	
The largest public opinion index of the first week	y_1	y_2	y_3	
The smallest public opinion index of the first week	z_1	z_2	z_3	
.....				
Nth Week				

[0051] S50: Obtain original sample data based on the first portrait data and the medical data.

[0052] Specifically, the first portrait data is used as the feature data for model training, and the medical data is used as the label data for model training, so as to obtain the original sample data.

[0053] S60: Clean the original sample data to obtain sample data to be processed.

[0054] Specifically, since the original sample data may include missing values or outlier, in order to further ensure the accuracy of subsequent model predictions, it is necessary to perform data cleaning on the original sample data to ensure the quality of the sample data to be processed.

[0055] S70: perform lag processing on the sample data to be processed to obtain lagged sample data.

[0056] Wherein, the lag processing is a feature engineering method, by expanding the sample data set, that is, increasing the feature portrait to collect more information. From a business logic perspective, that is the effect of the delay feature. Specifically, due to different themes predicted by some models, the corresponding sample data has hysteresis, such as disease outbreaks or economic-related data. In this embodiment, it is assumed that the prediction theme is to predict chickenpox, and there is a hysteresis in the outbreak of chickenpox. For example, if the temperature suddenly rises this week and the climate is humid, it may not cause the outbreak of chickenpox this week,

but it will usher in the next week. Therefore, it is necessary to perform lag processing on the sample data to be processed to ensure the accuracy of subsequent model predictions. Specifically, the sample data to be processed is subjected to n times lag processing (n generally takes 1~3). Assuming that n takes 1, the sample data to be processed is subjected to lag processing, that is, the original data of the first week is regarded as the data of the second week, the data of the second week is regarded as the data of the third week, and so on to obtain the lagged sample data. Assuming that n takes 2, because it is performed again on the basis of the sample data obtained the first time, the sample data to be processed is subjected to lag processing, that is, the original data of the first week is regarded as the data of the third week, the data of the second week is regarded as the data of the fourth week, and so on to obtain the lagging data. Integrate the lagging data obtained each time to obtain lagged sample data to achieve the purpose of expanding the sample data set.

[0057] Finally, the concat function is used to combine the lagged sample data obtained by multiple lag processing with the sample data to be processed into a data frame, that is, the lagged sample data. Wherein, the concat function is a function used to connect two or more arrays. The data frame is a two-dimensional data structure, that is, the data is arranged in a table of rows and columns.

[0058] S80: Perform feature expansion processing on the lagged sample data to obtain target sample data.

[0059] Specifically, in order to expand the sample data set and further improve the accuracy of model prediction, in this embodiment, feature expansion processing is performed on the lagged sample data to obtain target sample data, so as to achieve the purpose of further expanding the sample data set.

[0060] S90: Use an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model. The improved multi-granularity cascaded random forest algorithm includes a pooling layer that is used to retain data characteristics.

[0061] Wherein, the improved multi-granularity cascaded random forest algorithm is an algorithm that introduces the idea of pooling in the convolutional neural network into the multi-granularity cascaded random forest algorithm. The multi-granularity

cascaded random forest algorithm is a decision tree ensemble method that stacks multiple layers of random forests in a cascaded manner to obtain better feature representation and learning performance. The algorithm does not need to adjust hyperparameters too much to achieve good results performance.

[0062] Wherein, each layer in the multi-granularity cascaded random forest (Gcforest) is composed of multiple random forests. The feature information of the input feature vector is learned through random forest, and then input to the next layer after processing. In order to enhance the generalization ability of the model, multiple different types of random forests are selected for each layer. For example, two random forest structures are selected for each layer, which are completely-random tree forests and random forests.

[0063] In this embodiment, first, according to preset keywords, a crawler tool is used to crawl public opinion data obtained by third-party information platforms, so as to determine at least one hit entry that is truly related to the predicted topic based on the public opinion data, so as to ensure the validity and accuracy of the subsequently obtained public opinion factors. Then obtain the public opinion index and medical data corresponding to the hit entry per unit time. Finally, the public opinion factor and the public opinion index with the time label are used as the original sample data, so that the model analyzes the public opinion data in a unit time of 20 years in history. Then, by performing data cleaning on the original sample data, the sample data to be processed is obtained to ensure the quality of the sample data to be processed. Then, lag processing is performed on the sample data to be processed to obtain lagged sample data to expand the sample data set. In addition, for data with hysteresis, the effect of delay feature can be realized to ensure the accuracy of model prediction. Then, feature expansion processing is performed on the lagged sample data to obtain the target sample data to achieve the purpose of further expanding the sample data set and improve the accuracy of model prediction. Finally, the improved multi-granularity cascaded random forest algorithm is used to train the target sample data to obtain the target prediction model to obtain better feature representation and learning performance, and the algorithm can achieve good performance without excessive adjustment of hyperparameters, ensure the accuracy of model predictions. In addition,

the improved multi-granularity cascaded random forest algorithm also includes a pooling layer to fully retain data characteristics and further improve the accuracy of model prediction.

[0064] In an embodiment, before step S10, the intelligent data analysis method further includes:

[0065] S101: Obtain meteorological factors and corresponding meteorological data.

[0066] Understandably, in this embodiment, different portrait data can be selected according to different predicted themes. In this embodiment, the prediction of chickenpox is taken as an example for illustration. Because the climate conditions and the chickenpox virus are very closely related, the daily meteorological factors of 20 years in different regions were selected as part of the portrait data. The meteorological factors include, but are not limited to, day and night temperature, day and night pressure, day and night precipitation, humidity, illumination intensity, and wind in different regions.

[0067] S102: Take the meteorological factors and the corresponding meteorological data as second portrait data.

[0068] Wherein, the second portrait data refers to the feature data that uses meteorological factors and corresponding meteorological data as model training. Specifically, the method of establishing portrait data for meteorological factors is consistent with step S40, that is, the meteorological factors can be used as column labels, and the meteorological conditions of the Nth week are used as row labels to establish the second portrait data. Wherein, the meteorological conditions in the Nth week include but are not limited to the average meteorological conditions in the Nth week (such as average precipitation), the maximum meteorological conditions in the Nth week (such as maximum precipitation), and the minimum meteorological conditions in the Nth week (such as minimum precipitation).

[0069] Correspondingly, in step S50, that is, obtaining the original sample data based on the first portrait data and medical data includes:

[0070] S51: Take the first profile data, second profile data and medical data as the original sample data.

[0071] In this embodiment, the meteorological situation is combined with the idea of

mass dissemination of public opinion data to effectively predict the period of disease outbreak and improve the accuracy of model prediction.

[0072] In an embodiment, as shown in FIG. 3, in step S60, data cleaning is performed on the original sample data to obtain sample data to be processed, which specifically includes the following steps:

[0073] S61: Fill missing value of the original sample data to obtain first sample data.

[0074] Wherein, the missing value filling methods include, but are not limited to, mean filling, mode filling, median filling, expectation maximization method, multiple filling and k-means clustering method. Specifically, taking the k-means clustering method for filling as an example, the portrait data where the missing value is located is clustered, and the missing value is filled with the mean value of the clustered cluster.

[0075] S62: Perform outlier detection on the first sample data to obtain at least one outlier and mark the outlier as null.

[0076] S63: Fill missing value of the outlier marked as null to obtain the sample data to be processed.

[0077] Specifically, outlier detection includes, but is not limited to, the use of statistical variable analysis (such as box plot analysis, mean, maximum and minimum analysis, and the 3σ rule), distance-based methods, density-based outlier detection, and isolation forest, etc. In this embodiment, take the 3σ rule as an example. If the data obeys the normal distribution, under the 3σ principle, an outlier is defined as a value that deviates from the average value by more than 3 times the standard deviation in a set of measured values. Because under the assumption of normal distribution, the probability of occurrence of values beyond 3σ from the average value is less than 0.003, that is, data exceeding $\mu + 3\sigma$ and data not exceeding $\mu - 3\sigma$ are regarded as outliers.

[0078] Specifically, because the sample data corresponding to the outlier is not necessarily unnecessary, if the sample data corresponding to the outlier is deleted directly, it will cause the lack of features in the sample data, affect the quality of the sample data, and then affect the accuracy of the model prediction. Therefore, in this embodiment, the outliers are deleted and marked as null values, and then filling

missing values is performed again for the outliers marked as null values to obtain the sample data to be processed. In this embodiment, the sample data to be processed is obtained by filling missing values for the outliers marked as null, so as to avoid the problem of directly removing the sample data corresponding to the outliers, resulting in the lack of this part of the sample data, and affecting the accuracy of the model prediction.

[0079] In this embodiment, the first sample data is obtained by filling missing values of the original sample data. Then outlier detection is performed on the first sample data to obtain at least one outlier. By processing the outliers and missing values in the sample data, the purpose of data cleaning is achieved, and the quality of the sample data is guaranteed. Then, the obtained outliers are marked as null, so that the outliers marked as null can be processed by filling missing values again to obtain samples to be processed. In order to ensure the quality and standardization of the sample data and improve the accuracy of model prediction, fill missing values twice for the original sample data.

[0080] In one embodiment, as shown in FIG. 4, in step S80, that is, performing feature expansion processing on lagged sample data to obtain target sample data, specifically includes the following steps:

[0081] S81: Perform feature expansion on the lagged sample data to obtain a feature value corresponding to at least one statistical indicator.

[0082] S82: Splice the feature value with the lagged sample data to obtain the target sample data.

[0083] Wherein, the statistical indicators include, but are not limited to, the maximum, minimum, mean and standard deviation corresponding to each row of data. Each statistical indicator added to the lagged sample data as a new column, to expand the data set and increase the feature profile to collect more feature information and improve the accuracy of model prediction. Understandably, the lagged sample data is a matrix, and the eigenvalues and the lagged sample data are spliced together to obtain the target sample data, that is, N columns are added to the sample matrix, and N is the number of the statistical indicators (such as the maximum value, minimum value and average value of the data corresponding to each row). The maximum, minimum and

average values of the data corresponding to each row are the characteristic values.

[0084] In this embodiment, the feature value corresponding to at least one statistical indicator is obtained by performing feature expansion on the lagged sample data. The feature value and the lagged sample data are spliced together to obtain the target sample data to expand the data set and increase the feature portrait to collect more feature information and improve the accuracy of model prediction.

[0085] In an embodiment, as shown in FIG. 5, after step S80, the intelligent data analysis method further includes the following steps:

[0086] S111: Perform variance analysis on the target sample data and remove data of which variance is less than a preset variance threshold, to obtain the second sample data.

[0087] S112: Perform singular value decomposition on the second sample data to update the target sample data.

[0088] Specifically, because the data size being too large is as bad as being too small, huge data size in data analysis applications will actually produce worse performance. Therefore, it is necessary to filter the target sample data to remove redundant data, so as to reduce the number of data columns while ensuring that the missing data information is as little as possible.

[0089] Wherein, the variance analysis refers to the analysis based on the variance of the data column to remove the sequence with too small variance (that is, less than the preset variance threshold) to obtain the second sample data. Specifically, the size of the variance describes the amount of information of a variable, and the sequence with too small variance is considered to contain less information. Therefore, all data columns with small variances are removed to achieve the effect of data dimensionality reduction, reduce the amount of data processing, and improve the efficiency of subsequent model training.

[0090] Specifically, the target sample data contains multiple features, but some features have little effect on the prediction accuracy of the model, or it can be considered that features that are too correlated can be replaced equally. Therefore, redundant variables can be removed to achieve the purpose of data dimensionality reduction and save model training time. Specifically, when the variance analysis is used, the data column

of which variance is less than the preset variance threshold is removed, so the accuracy of the variance analysis depends on the preset variance threshold. Therefore, in order to further remove redundant data and ensure that the missing data information is as little as possible, it is also necessary to perform singular value decomposition on the second sample data to remove redundant data to achieve the purpose of data compression and ensure the quality of the target sample data.

[0091] In this embodiment, by performing variance analysis on the target sample data, the data with variance less than the preset variance threshold is removed to obtain the second sample data, thus redundant data can be removed to reduce the number of data columns while ensuring that the missing data information is as little as possible, thereby saving model training time. Then, singular value decomposition is performed on the second sample data, and the target sample data is updated to further remove redundant data and ensure the quality of the target sample data.

[0092] In one embodiment, the improved multi-granularity cascaded random forest algorithm includes multi-particle scanning algorithm and cascaded random forest algorithm, and the multi-particle scanning algorithm corresponds to at least one sliding window, as shown in FIG. 6, in step S90, specifically including the following steps:

[0093] S91: Use the multi-particle scanning algorithm to perform multi-particle scanning on target sample data according to at least one sliding window to obtain at least one intermediate data.

[0094] Wherein, multi-particle scanning refers to scanning the target sample data using a sliding window to obtain at least one intermediate data. In this embodiment, sliding windows of different dimensions can be set. Understandably, the sliding window can be an $i*j$ window. For example, the sample data row label of the target table is the i -th week, and the window size of sliding window can be 2 (every 2 weeks), 4 (every month), 12 (every quarter), etc. It should be noted that the sliding window can scan at least one feature portrait, that is, every row, every two rows, and every j row can be scanned to maximize the search for the internal relationship between the feature and the label set, and the feature and the feature.

[0095] S92: Based on the pooling layer, perform pooling processing on at least one intermediate data, to obtain data to be trained.

[0096] In this embodiment, at least one intermediate data is pooled through the pooling layer to obtain the data to be trained, so as to achieve the purpose of reducing the dimensionality of the data, reduce the amount of calculation, and improve the efficiency of model training.

[0097] S93: Use the cascaded random forest algorithm to train the data to be trained, to obtain the target prediction model.

[0098] Specifically, the multi-granularity cascaded random forest algorithm is based on the idea of neural network integration, and uses the label column $cforest_i$ predicted by the i -th complete-random tree forest and the label column $rforest_i$ predicted by random forest as the portrait column that is continuously added to the target sample data to achieve feature expansion. And finally obtain the following feature portrait $[orgf_1, orgf_2, \dots, orgf_n, cforest_1, rforest_1, \dots, cforest_k, rforest_k]$. Wherein, $orgf$ is the target sample data. Finally, input the feature portrait into the last m (m generally takes 3~5, general order of magnitude takes 3, tens of millions order of magnitude takes 3~4, more than tens of millions of order takes 4~5) random forest for prediction, and take the final Max value as the final predicted probability value.

[0099] Specifically, the obtained data to be trained is input into the cascade forest for training. For example, in this embodiment, sliding windows of three dimensions are used. First, the sliding window of the first dimension is used to scan to obtain a feature vector, and then the original feature vector is input into the complete-random tree forest and random forest to obtain two prediction sequences (i.e. $cforest_i$ and $rforest_i$). Then the two prediction sequences are spliced together to obtain the first feature vector, and the original feature vector is input into the first layer of cascaded forest for training to obtain the first prediction sequence. Then the obtained first prediction sequence is spliced with the first feature vector to obtain the second feature vector, which is used as the input data of the second layer of cascaded forest; the second prediction sequence obtained by training of the second layer of cascaded forest is then spliced with the third feature vector obtained by the sliding window of the second dimension (same method as the first feature vector) and used as the input

data of the third layer of cascaded forest; the third prediction sequence obtained by training of the third layer of cascaded forest is then spliced with the fourth feature vector obtained by the sliding window of the third dimension and used as the input of the next layer. The above process is repeated until convergence, and the target prediction model is obtained.

[00100] In this embodiment, the multi-particle scanning algorithm is used to perform multi-particle scanning on the target sample data according to at least one sliding window to obtain at least one intermediate data, so as to maximize the search for the internal correlation between features and label sets, and features and features. Then, by combining the pooling layer, at least one intermediate data is pooled to obtain the data to be trained, so as to combine machine learning and neural network ideas to obtain more intuitive and inaccessible information to enrich the model and further improve the accuracy of model prediction.

[00101] In an embodiment, as shown in FIG. 7, in step S92, that is, based on the pooling layer, pooling is performed on at least one intermediate data to obtain the data to be trained, which specifically includes the following steps:

[00102] S921: Select two adjacent intermediate data as a data group to be processed to obtain at least one data group to be processed corresponding to the intermediate data.

[00103] S922: Perform an averaging operation on each data group to be processed to obtain a first data sequence.

[00104] S923: Perform a minimum value operation on each data group to be processed to obtain a second data sequence. The second data sequence includes the minimum value of the two-intermediate data of each group to be processed.

[00105] S924: Perform a maximum value operation on each data group to be processed to obtain a third data sequence, and the third data column includes sequence includes the maximum value of the two-intermediate data of each group to be processed.

[00106] S925: Splice the first data sequence, the second data sequence, and the third data sequence to obtain data to be trained.

[00107] Specifically, from a business logic level, model prediction requires more linear or non-linear methods to spatially distort data, so as to obtain more information

that is not intuitively available to enrich the model. Therefore, in this embodiment, the three pooling methods are used for pooling at least one intermediate data, and then the results obtained by pooling in each method are integrated to obtain the data to be trained, so as to obtain more intuitive and inaccessible information to enrich the model and fully retain data characteristics. Assuming that a certain column of portrait data in the intermediate data is $Feature : f_1, f_2, f_3, f_4, f_5, \dots, f_n$, the following three pooling methods are used to pool at least one intermediate data.

$$Feature_new_1: (f_1 + f_2) / 2, (f_2 + f_3) / 2, \dots, (f_{n-1} + f_n) / 2$$

$$Feature_new_2: \max(f_1, f_2), \max(f_2, f_3), \dots, \max(f_{n-1}, f_n)$$

$$[00108] \quad Feature_new_3: \min(f_1, f_2), \min(f_2, f_3), \dots, \min(f_{n-1}, f_n)$$

[00109] In this embodiment, at least one intermediate data is pooled by adopting three pooling methods, and then the results obtained by pooling in each method are integrated to obtain the data to be trained, so as to fully retain the data characteristics ensure the sample data quality, and improve the accuracy of model prediction.

[00110] It should be understood that the size of the sequence number of each step in the foregoing embodiment does not mean the order of execution, and the execution sequence of each process should be determined by its function and internal logic, and should not constitute any limitation to the implementation process of the embodiment of the present application.

[00111] In one embodiment, an intelligent data analysis apparatus is provided, and the intelligent data analysis apparatus corresponds one-to-one with the intelligent data analysis method in the foregoing embodiment. As shown in FIG. 8, the intelligent data analysis apparatus includes a public opinion data acquisition module 10, a hit entry determination module 20, a public opinion index acquisition module 30, a first portrait data acquisition module 40, an original sample data acquisition module 50, and a to-be-processed sample data acquisition module 60, a lagged sample data acquisition module 70, a target sample data acquisition module 80, and a target prediction model acquisition module 90. The detailed description of each functional module is as follows:

[00112] The public opinion data acquisition module 10 is provided for using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword.

[00113] The hit entry determination module 20 is provided for determining at least one hit entry based on the public opinion data. The hit entry corresponds to a public opinion factor.

[00114] The public opinion index acquisition module 30 is provided for obtaining medical data within a historical unit time and a public opinion index corresponding to the hit entry. The public opinion index carries a time label.

[00115] The first portrait data acquisition module 40 is provided for taking the public opinion factor and the public opinion index with the time label as first portrait data.

[00116] The original sample data acquisition module 50 is provided for obtaining original sample data based on the first portrait data and the medical data.

[00117] The to-be-processed sample data acquisition module 60 is provided for cleaning the original sample data to obtain the sample data to be processed.

[00118] The lagged sample data acquisition module 70 is provided for performing lag processing on the sample data to be processed to obtain lagged sample data.

[00119] The target prediction model acquisition module 80 is provided for performing feature expansion processing on the lagged sample data to obtain target sample.

[00120] The data acquisition module of the target prediction model 90 is provided for using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model. The improved multi-granularity cascaded random forest algorithm comprises a pooling layer that is used to retain data characteristics.

[00121] Specifically, the to-be-processed sample data acquisition module includes a first sample data acquisition unit, an outlier acquisition unit, and a to-be-processed sample data acquisition unit.

[00122] The first sample data acquisition unit is provided for filling missing value of the original sample data to obtain first sample data.

[00123] The outlier acquisition unit is provided for performing outlier detection on the first sample data to obtain at least one outlier and marking the outlier as null.

[00124] The to-be-processed sample data acquisition unit is provided for filling missing value of the outlier marked as null to obtain sample data to be processed.

[00125] Specifically, the target sample data acquisition module includes a feature value acquisition unit and a target sample data acquisition unit.

[00126] The feature value obtaining unit is provided for performing feature expansion on the lagged sample data to obtain a feature value corresponding to at least one statistical indicator.

[00127] The target sample data acquisition unit is provided for splicing the feature value with the lagged sample data to obtain the target sample data.

[00128] Specifically, the intelligent data analysis apparatus includes a second sample data acquisition unit and a target sample data update unit.

[00129] The second sample data acquisition unit is provided for performing variance analysis on the target sample data and removing data of which variance is less than a preset variance threshold, to obtain second sample data.

[00130] The target sample data update unit is provided for performing singular value decomposition on the second sample data to update the target sample data.

[00131] Specifically, the improved multi-granularity cascaded random forest algorithm includes a multi-particle scanning algorithm and a cascaded random forest algorithm. The multi-particle scanning algorithm corresponds to at least one sliding window. The target prediction model acquisition module includes a target prediction model, a to-be-trained data acquisition unit, and a target prediction model acquisition unit.

[00132] The intermediate data acquisition unit is provided for using the multi-particle scanning algorithm to perform multi-particle scanning on the target sample data according to at least one of the sliding windows, to obtain at least one intermediate data.

[00133] The to-be-trained data obtaining unit is provided for based on the pooling layer, performing pooling processing on at least one of the intermediate data to obtain data to be trained.

[00134] The target prediction model acquisition unit is provided for using the cascaded random forest algorithm to train the data to be trained, to obtain the target prediction model.

[00135] Specifically, the to-be-trained data obtaining unit includes a to-be-processed data group acquisition subunit, a first data sequence acquisition subunit, a second data

sequence acquisition subunit, a third data sequence acquisition subunit, and a to-be-trained data acquisition subunit.

[00136] The to-be-processed data group acquisition subunit is provided for selecting two adjacent intermediate data as a data group to be processed to obtain at least one data group to be processed corresponding to the intermediate data.

[00137] The first data sequence acquisition subunit is provided for performing an averaging operation on each data group to be processed to obtain a first data sequence.

[00138] The second data sequence acquisition subunit is provided for performing a minimum operation on each data group to be processed to obtain a second data sequence, the second data sequence comprising a minimum value of two intermediate data of each data group to be processed.

[00139] The third data sequence acquisition subunit is provided for performing a maximum operation on each data group to be processed to obtain a third data sequence, the third data sequence comprising a maximum value of two intermediate data of each data group to be processed.

[00140] The to-be-trained data acquisition subunit is provided for splicing the first data sequence, the second data sequence and the third data sequence to obtain data to be trained.

[00141] For the specific definition of the intelligent data analysis apparatus, please refer to the above definition of the intelligent data analysis method, which will not be repeated here. Each module in the above-mentioned intelligent data analysis apparatus can be implemented in whole or in part by software, hardware and a combination thereof. The foregoing modules may be embedded in the form of hardware or independent of the processor in the computer device or may be stored in the memory of the computer device in the form of software, so that the processor can call and execute the operations corresponding to the foregoing modules.

[00142] In one embodiment, a computer device is provided. The computer device may be a server, and its internal structure diagram may be as shown in FIG. 10. The computer device includes a processor, a memory, a network interface and a database connected by a system bus. Wherein, the processor of the computer device is used to provide calculation and control capabilities. The memory of the computer device

includes a readable storage medium and an internal memory. The readable storage medium stores an operating system, computer readable instructions, and a database. The internal memory provides an environment for the operation of the operating system and computer readable instructions in the readable storage medium. The database of the computer device is used to store data generated or obtained during the execution of the intelligent data analysis method, such as target sample data. The network interface of the computer device is used to communicate with an external terminal through a network connection. The computer readable instructions are executed by the processor to realize an intelligent data analysis method.

[00143] In one embodiment, a computer device is provided, including a memory, a processor, and computer-readable instructions stored in the memory and executable on the processor. The processor executes the computer-readable instructions to implement the steps of the intelligent data analysis method, for example, steps S10-S90 shown in FIG. 2 or the steps shown in FIG. 3 to FIG. 7. Alternatively, the processor implements the functions of the modules/units in this embodiment of the intelligent data analysis apparatus when the processor executes the computer-readable instructions, such as the functions of the modules/units shown in FIG. 8. To avoid repetition, details are not repeated here.

[00144] In an embodiment, one or more readable storage media storing computer readable instructions are provided, and the computer readable storage medium stores computer readable instructions, wherein when the computer readable instructions are executed by one or more processors, the steps of the intelligent data analysis method in the foregoing embodiment are realized by the one or more processors, for example, steps S10-S90 shown in FIG. 2 or shown in FIGS. 3 to 7. To avoid repetition, details are not repeated here. Alternatively, when the computer-readable instructions are executed by the processor, the function of each module/unit in the embodiment of the above-mentioned intelligent data analysis apparatus is realized, for example, the function of each module/unit shown in FIG. To avoid repetition, details are not repeated here. The readable storage medium in this embodiment includes a nonvolatile readable storage medium and a volatile readable storage medium.

[00145] A person of ordinary skill in the art can understand that all or part of the

processes in the above-mentioned embodiment methods can be implemented by instructing relevant hardware through a computer program. The computer program can be stored in a non-volatile computer readable storage. In the medium, when the computer program is executed, it may include the procedures of the embodiments of the above-mentioned method. Wherein, any reference to memory, storage, database or other media used in the embodiments provided in the present application may include non-volatile and/or volatile memory. Non-volatile memory may include read-only memory (ROM), programmable ROM (PROM), electrically programmable ROM (EPROM), electrically erasable programmable ROM (EEPROM), or flash memory. Volatile memory may include random access memory (RAM) or external cache memory. As an illustration and not a limitation, RAM is available in many forms, such as static RAM (SRAM), dynamic RAM (DRAM), synchronous DRAM (SDRAM), double data rate SDRAM (DDRSDRAM), enhanced SDRAM (ESDRAM), synchronous link (Synchlink) DRAM (SLDRAM), Rambus direct RAM (RDRAM), direct Rambus dynamic RAM (DRDRAM), and Rambus dynamic RAM (RDRAM), etc.

[00146] A person skilled in the art can clearly understand that, for the convenience and conciseness of description, only the division of the above functional units and modules is used as an example. In actual applications, the above functions can be allocated to different functional units or modules to complete. That is, the internal structure of the device is divided into different functional units or modules to complete all or part of the functions described above.

[00147] The above-mentioned embodiments are only used to illustrate the technical solutions of the present application, but not to limit them; although the embodiments of the present application are described in detail with reference to the foregoing embodiments, those of ordinary skill in the art should understand that it is still possible to modify the technical solutions recorded in the foregoing embodiments, or equivalently replace some of the technical features; and these modifications or replacements do not cause the essence of the corresponding technical solutions to deviate from the spirit and scope of the technical solutions of the embodiments of the present application, and should be included in the protection scope of the embodiments

of the present application.

What is claimed is:

1. An intelligent data analysis method, comprising:
 - using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword;
 - determining at least one hit entry based on the public opinion data; and the hit entry corresponding to a public opinion factor;
 - obtaining medical data within a historical unit time and a public opinion index corresponding to the hit entry; and the public opinion index carrying a time label;
 - taking the public opinion factor and the public opinion index with the time label as first portrait data;
 - obtaining original sample data based on the first portrait data and the medical data;
 - cleaning the original sample data to obtain sample data to be processed;
 - performing lag processing on the sample data to be processed to obtain lagged sample data;
 - performing feature expansion processing on the lagged sample data to obtain target sample data; and
 - using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model; and the improved multi-granularity cascaded random forest algorithm comprising a pooling layer that is used to retain data characteristics.
2. The intelligent data analysis method according to claim 1, wherein before the step of using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword, the intelligent data analysis method further comprises:
 - obtaining meteorological factors and corresponding meteorological data; and
 - taking the meteorological factors and the corresponding meteorological data as second portrait data;the step of obtaining original sample data based on the first portrait data and the medical data comprises:
 - taking the first portrait data, the second portrait data and the medical data as original

sample data.

3. The intelligent data analysis method according to claim 1, wherein the step of cleaning the original sample data to obtain sample data to be processed, comprises:
filling missing value of the original sample data to obtain first sample data;
performing outlier detection on the first sample data to obtain at least one outlier, and
marking the outlier as null; and
filling missing value of the outlier marked as null to obtain sample data to be processed.
4. The intelligent data analysis method according to claim 1, wherein the step of performing lag processing on the sample data to be processed to obtain lagged sample data, comprises:
performing feature expansion on the lagged sample data to obtain a feature value corresponding to at least one statistical indicator; and
splicing the feature value with the lagged sample data to obtain the target sample data.
5. The intelligent data analysis method according to claim 1 or 4, wherein after the step of obtaining the target sample data, the intelligent data analysis method further comprises:
performing variance analysis on the target sample data, and removing data of which variance is less than a preset variance threshold, to obtain second sample data; and
performing singular value decomposition on the second sample data to update the target sample data.
6. The intelligent data analysis method according to claim 1, wherein the improved multi-granularity cascaded random forest algorithm comprises a multi-particle scanning algorithm and a cascaded random forest algorithm, the multi-particle scanning algorithm corresponds to at least one sliding window;
the step of using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model comprises:
using the multi-particle scanning algorithm to perform multi-particle scanning on the target sample data according to at least one of the sliding windows, to obtain at least one intermediate data;
based on the pooling layer, performing pooling processing on at least one of the

intermediate data to obtain data to be trained;

using the cascaded random forest algorithm to train the data to be trained, to obtain the target prediction model.

7. The intelligent data analysis method according to claim 6, wherein the step of performing pooling processing on at least one of the intermediate data to obtain data to be trained, comprises:

selecting two adjacent intermediate data as a data group to be processed to obtain at least one data group to be processed corresponding to the intermediate data;

performing an averaging operation on each data group to be processed to obtain a first data sequence;

performing a minimum operation on each data group to be processed to obtain a second data sequence, the second data sequence comprising a minimum value of two intermediate data of each data group to be processed;

performing a maximum operation on each data group to be processed to obtain a third data sequence, the third data sequence comprising a maximum value of two intermediate data of each data group to be processed; and

splicing the first data sequence, the second data sequence and the third data sequence to obtain data to be trained.

8. An intelligent data analysis apparatus, comprising:

a public opinion data acquisition module, for using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword;

a hit entry determination module, for determining at least one hit entry based on the public opinion data; and the hit entry corresponding to a public opinion factor;

a public opinion index acquisition module, for obtaining medical data within a historical unit time and a public opinion index corresponding to the hit entry; and the public opinion index carrying a time label;

a first portrait data acquisition module, for taking the public opinion factor and the public opinion index with the time label as first portrait data;

an original sample data acquisition module, for obtaining original sample data based on the first portrait data and the medical data;

a to-be-processed sample data acquisition module, for cleaning the original sample data to obtain sample data to be processed;

a lagged sample data acquisition module, for performing lag processing on the sample data to be processed to obtain lagged sample data;

a target sample data acquisition module, for performing feature expansion processing on the lagged sample data to obtain target sample data;

a target prediction model acquisition module, for using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model; and the improved multi-granularity cascaded random forest algorithm comprising a pooling layer that is used to retain data characteristics.

9. A computer device, comprises a memory, a processor, and computer readable instructions stored in the memory and run in the processor, wherein the computer readable instructions are executed by the processor to perform the steps of:
 - using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword;
 - determining at least one hit entry based on the public opinion data; and the hit entry corresponding to a public opinion factor;
 - obtaining medical data within a historical unit time and a public opinion index corresponding to the hit entry; and the public opinion index carrying a time label;
 - taking the public opinion factor and the public opinion index with the time label as first portrait data;
 - obtaining original sample data based on the first portrait data and the medical data;
 - cleaning the original sample data to obtain sample data to be processed;
 - performing lag processing on the sample data to be processed to obtain lagged sample data;
 - performing feature expansion processing on the lagged sample data to obtain target sample data; and
 - using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model; and the improved multi-granularity cascaded random forest algorithm comprising a pooling layer that is used to retain data characteristics.

10. The computer device according to claim 9, wherein before the step of using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword, the intelligent data analysis method further comprises:
obtaining meteorological factors and corresponding meteorological data; and
taking the meteorological factors and the corresponding meteorological data as second portrait data;
the step of obtaining original sample data based on the first portrait data and the medical data comprises:
taking the first portrait data, the second portrait data and the medical data as original sample data.
11. The computer device according to claim 9, wherein the step of cleaning the original sample data to obtain sample data to be processed, comprises:
filling missing value of the original sample data to obtain first sample data;
performing outlier detection on the first sample data to obtain at least one outlier, and
marking the outlier as null; and
filling missing value of the outlier marked as null to obtain sample data to be processed.
12. The computer device according to claim 9, wherein the step of performing lag processing on the sample data to be processed to obtain lagged sample data, comprises:
performing feature expansion on the lagged sample data to obtain a feature value corresponding to at least one statistical indicator; and
splicing the feature value with the lagged sample data to obtain the target sample data.
13. The computer device according to claim 9, wherein the improved multi-granularity cascaded random forest algorithm comprises a multi-particle scanning algorithm and a cascaded random forest algorithm, the multi-particle scanning algorithm corresponds to at least one sliding window;
the step of using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model comprises:
using the multi-particle scanning algorithm to perform multi-particle scanning on the target sample data according to at least one of the sliding windows, to obtain at least

one intermediate data;

based on the pooling layer, performing pooling processing on at least one of the intermediate data to obtain data to be trained;

using the cascaded random forest algorithm to train the data to be trained, to obtain the target prediction model.

14. The computer device according to claim 13, wherein the step of performing pooling processing on at least one of the intermediate data to obtain data to be trained, comprises:

selecting two adjacent intermediate data as a data group to be processed to obtain at least one data group to be processed corresponding to the intermediate data;

performing an averaging operation on each data group to be processed to obtain a first data sequence;

performing a minimum operation on each data group to be processed to obtain a second data sequence, the second data sequence comprising a minimum value of two intermediate data of each data group to be processed;

performing a maximum operation on each data group to be processed to obtain a third data sequence, the third data sequence comprising a maximum value of two intermediate data of each data group to be processed; and

splicing the first data sequence, the second data sequence and the third data sequence to obtain data to be trained.

15. One or more readable storage media storing computer readable instructions, wherein when the computer readable instructions executed by one or more processors, one or more processors are caused to implement the steps of:

using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword;

determining at least one hit entry based on the public opinion data; and the hit entry corresponding to a public opinion factor;

obtaining medical data within a historical unit time and a public opinion index corresponding to the hit entry; and the public opinion index carrying a time label;

taking the public opinion factor and the public opinion index with the time label as first portrait data;

obtaining original sample data based on the first portrait data and the medical data;
cleaning the original sample data to obtain sample data to be processed;
performing lag processing on the sample data to be processed to obtain lagged sample data;
performing feature expansion processing on the lagged sample data to obtain target sample data; and
using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model; and the improved multi-granularity cascaded random forest algorithm comprising a pooling layer that is used to retain data characteristics.

16. The readable storage media according to claim 15, wherein before the step of using a crawler tool to crawl public opinion data obtained by a third-party information platform according to a preset keyword, the intelligent data analysis method further comprises:

obtaining meteorological factors and corresponding meteorological data; and
taking the meteorological factors and the corresponding meteorological data as second portrait data;
the step of obtaining original sample data based on the first portrait data and the medical data comprises:
taking the first portrait data, the second portrait data and the medical data as original sample data.

17. The readable storage media according to claim 15, wherein the step of cleaning the original sample data to obtain sample data to be processed, comprises:

filling missing value of the original sample data to obtain first sample data;
performing outlier detection on the first sample data to obtain at least one outlier, and marking the outlier as null; and
filling missing value of the outlier marked as null to obtain sample data to be processed.

18. The readable storage media according to claim 15, wherein the step of performing lag processing on the sample data to be processed to obtain lagged sample data, comprises:

- performing feature expansion on the lagged sample data to obtain a feature value corresponding to at least one statistical indicator; and
- splicing the feature value with the lagged sample data to obtain the target sample data.
19. The readable storage media according to claim 15, wherein the improved multi-granularity cascaded random forest algorithm comprises a multi-particle scanning algorithm and a cascaded random forest algorithm, the multi-particle scanning algorithm corresponds to at least one sliding window;
- the step of using an improved multi-granularity cascaded random forest algorithm to train the target sample data to obtain a target prediction model comprises:
- using the multi-particle scanning algorithm to perform multi-particle scanning on the target sample data according to at least one of the sliding windows, to obtain at least one intermediate data;
- based on the pooling layer, performing pooling processing on at least one of the intermediate data to obtain data to be trained;
- using the cascaded random forest algorithm to train the data to be trained, to obtain the target prediction model.
20. The readable storage media according to claim 19, wherein the step of performing pooling processing on at least one of the intermediate data to obtain data to be trained, comprises:
- selecting two adjacent intermediate data as a data group to be processed to obtain at least one data group to be processed corresponding to the intermediate data;
- performing an averaging operation on each data group to be processed to obtain a first data sequence;
- performing a minimum operation on each data group to be processed to obtain a second data sequence, the second data sequence comprising a minimum value of two intermediate data of each data group to be processed;
- performing a maximum operation on each data group to be processed to obtain a third data sequence, the third data sequence comprising a maximum value of two intermediate data of each data group to be processed; and
- splicing the first data sequence, the second data sequence and the third data sequence to obtain data to be trained.