

CLAIMS

What is claimed is:

1. A method for detecting disease probability, applied to an insurance system, the method comprising:

collecting each datum associated with a user, and performing feature analysis on each collected datum to determine a feature type of each datum, wherein each datum comprises behavior information and health information of the user;

in response to a determination that the datum is a missing value datum, performing mean imputation or multiple imputation to the missing value datum;

in response to a determination that the datum is an outlier datum, screening the outlier datum, and screening out the datum whose outlier is less than a preset threshold, and treating the screened datum as the missing value datum;

constructing a multi-dimensional data set according to each datum after being feature processed;

performing random sampling on the multi-dimensional data set to divide into a test set and a training set;

building a model based on the training set to obtain a regression decision tree;

testing the regression decision tree according to the test set, to calculate the disease probability of the user.

2. The method of claim 1, wherein the mean imputation comprises: performing interpolation using an average value, or performing interpolation using mode.

3. The method of claim 1, wherein the step of constructing a multi-dimensional data set according to each datum after being feature processed comprises:

determining feature saturation corresponding to each datum after being feature processed;

screening each datum according to the feature saturation to screen out the datum whose feature saturation reaches a preset saturation;

constructing the multi-dimensional data set based on each screened datum.

4. The method of claim 1, wherein the step of testing the regression decision tree according to the test set, to calculate the disease probability of the user comprises:

inputting the data in the test set into the regression decision tree to obtain the numerical value corresponding to the number of trees in the regression decision tree;

calculating weighted average of each numerical value with a weight value of each tree in the regression decision tree to obtain a total value of the regression decision tree;

using the total value as the disease probability of the user.

5. A device for detecting disease probability, applied to an insurance system, the device comprising:

a processing module, configured for collecting each datum associated with a user, and performing feature analysis on each collected datum to determine a feature type of each datum; an interpolating unit, performing, in response to a determination that the datum is a missing value datum, mean imputation or multiple imputation to the missing value datum; and a screening unit, and screening, in response to a determination that the datum is an outlier datum, the outlier datum, and screening out the datum whose outlier is less than a preset threshold, and treating the screened datum as the missing value datum, wherein each datum comprises behavior information and health information of the user;

a constructing module, configured for constructing a multi-dimensional data set according to each datum after being feature processed;

a dividing module, configured for performing random sampling on the multi-dimensional data set to divide into a test set and a training set;

a building module, configured for building a model based on the training set to obtain a regression decision tree;

a calculating module, configured for testing the regression decision tree according to the test set, to calculate the disease probability of the user.

6. The device of claim 5, wherein the mean imputation comprises: performing interpolation using an average value, or performing interpolation using mode.

7. The device of claim 5, wherein the constructing module comprises:

a determining unit, configured for determining feature saturation corresponding to each datum being feature processed;

a screening unit, configured for screening each datum according to the feature saturation to screen out the datum whose feature saturation reaches a preset saturation;

a constructing unit, configured for constructing the multi-dimensional data set based on each screened datum.

8. The device of claim 5, wherein the calculating module comprises:

an inputting unit, configured for inputting the data in the test set into the regression decision tree to obtain the numerical value corresponding to the number of trees in the regression decision tree;

a calculating unit, configured for calculating weighted average of each numerical value with a weight value of each tree in the regression decision tree to obtain a total value of the regression decision tree;

a processing unit, configured for using the total value as the disease probability of the user.

9. An apparatus for detecting disease probability, applied to an insurance system, the apparatus comprising:

a processor,

a memory which stores a disease probability detecting program;

wherein the processor is configured for executing the disease probability detecting program to perform following operations:

collecting each datum associated with a user, and performing feature analysis on each collected datum to determine the feature type of each datum, wherein each datum comprises behavior information and health information of the user;

in response to a determination that the datum is a missing value datum, performing mean imputation or multiple imputation to the missing value datum;

in response to a determination that the datum is an outlier datum, screening the outlier datum, and screening out the datum whose outlier is less than a preset threshold, and treating the screened datum as the missing value datum;

constructing a multi-dimensional data set according to each datum after being feature processed;

performing random sampling on the multi-dimensional data set to divide into a test set and a training set;

building a model based on the training set to obtain a regression decision tree;

testing the regression decision tree according to the test set, to calculate the disease probability of the user.

10. The device of claim 9, wherein the mean imputation comprises: performing interpolation using an average value, or performing interpolation using mode.

11. The system of claim 9, wherein the processor is further configured for executing the disease probability detecting program to perform constructing a multi-dimensional data set according to each datum after being feature processed:

determining feature saturation corresponding to each datum being feature processed;

screening each datum according to the feature saturation to screen out the datum whose feature saturation reaches a preset saturation;

constructing the multi-dimensional data set based on each screened datum.

12. The system of claim 9, wherein the processor is further configured for executing the disease probability detecting program to perform testing the regression decision tree according to the test set, to calculate the disease probability of the user:

inputting the data in the test set into the regression decision tree to obtain the numerical value corresponding to the number of trees in the regression decision tree;

calculating weighted average of each numerical value with a weight value of each tree in the regression decision tree to obtain a total value of the regression decision tree;

using the total value as the disease probability of the user.

13. A computer-readable storage medium storing a disease probability detecting program, applied to an insurance system, which when executed by a processor performs following operations:

collecting each datum associated with a user, and performing feature analysis on each collected datum to determine the feature type of each datum, wherein each datum comprises behavior information and health information of the user;

in response to a determination that the datum is a missing value datum, performing mean imputation or multiple imputation to the missing value datum;

in response to a determination that the datum is an outlier datum, screening the outlier datum, and screening out the datum whose outlier is less than a preset threshold, and treating the screened datum as the missing value datum;

constructing a multi-dimensional data set according to each datum after being feature processed;

performing random sampling on the multi-dimensional data set to divide into a test set and a training set;

building a model based on the training set to obtain a regression decision tree;

testing the regression decision tree according to the test set, to calculate the disease probability of the user.

14. The computer-readable storage medium of claim 13, wherein the mean imputation comprises: performing interpolation using an average value, or performing interpolation using mode.

15. The computer-readable storage medium of claim 13, wherein when the disease probability detecting program executed by the processor, operations of constructing a multi-dimensional data set according to each datum after being feature processed are performed:

determining feature saturation corresponding to each datum after being feature processed;

screening each datum according to the feature saturation to screen out the datum whose feature saturation reaches a preset saturation;

constructing the multi-dimensional data set based on each screened datum.

16. The computer-readable storage medium of claim 13, wherein when the disease probability detecting program executed by the processor, operations of testing the regression

decision tree according to the test set, to calculate the disease probability of the user are performed:

- inputting the data in the test set into the regression decision tree to obtain the numerical value corresponding to the number of trees in the regression decision tree;

- calculating weighted average of each numerical value with a weight value of each tree in the regression decision tree to obtain a total value of the regression decision tree;

- using the total value as the disease probability of the user.