CLAIMS

1. A method for preventing model theft during classification, the method comprising:

obtaining feature information of a service object from a service requester through a predetermined or dynamically-generated interface;

inputting the feature information of the service object into a service classification model for initial prediction to obtain respective initial prediction probability corresponding to each category of a plurality of categories, wherein the service classification model includes a deep neural network including an input layer, a plurality of intermediate layers, and an output layer that are sequentially stacked;

determining, from the plurality of categories, a category corresponding to a maximum initial prediction probability as a target category;

determining disturbance data corresponding to each of one or more target layers based on the target category, wherein the one or more target layers is selected from the input layer, the plurality of intermediate layers, and the output layer, and wherein for each target layer, the disturbance data corresponding to the target layer is determined based on the target category and by performing:

determining a target function by using the data input into the target layer as an independent variable, wherein the target function is positively correlated with a prediction probability corresponding to the target category;

determining target data input into the target layer in the initial prediction;

calculating a gradient value of the target function by using the target data as a value of the independent variable; and

determining the disturbance data corresponding to the target layer based on the gradient value;

inputting the feature information into the classification model for subsequent prediction, wherein the subsequent prediction includes adding the disturbance data corresponding to each of the one or more target layers to data to be input into each of the one or more target layers, a

category from the plurality categories that corresponds to a maximum subsequent prediction probability obtained through the subsequent prediction remains the target category, and the maximum subsequent prediction probability is greater than the maximum initial prediction probability; and

outputting the target category and the maximum subsequent prediction probability to the service requester.

2. The method according to claim 1, wherein determining the disturbance data corresponding to the target layer based on the gradient value includes:

obtaining a constant and a function, wherein the function is not monotonically decreasing;

substituting the gradient value as the value of the independent variable into the function to obtain a target result; and

calculating a product of the constant and the target result as the disturbance data.

3. The method according to claim 2, wherein the function is a sign function.

4. The method according to claim 1, wherein at least one of the one or more target layers is one of the plurality of intermediate layers.

5. The method according to claim 1, wherein the service object includes at least one of a user, a merchant, a product, or an event.

6. A non-transitory computer readable medium storing contents that, when executed by one or more processors, cause the one or more processors to perform actions comprising:

obtaining feature information of a service object from a service requester through a predetermined or dynamically-generated interface;

inputting the feature information of the service object into a service classification model for initial prediction to obtain respective initial prediction probability corresponding to each

category of a plurality of categories, wherein the service classification model includes a deep neural network including an input layer, a plurality of intermediate layers, and an output layer that are sequentially stacked;

determining, from the plurality of categories, a category corresponding to a maximum initial prediction probability as a target category;

determining disturbance data corresponding to each of one or more target layers based on the target category, wherein the one or more target layers is selected from the input layer, the plurality of intermediate layers, and the output layer, and wherein for each target layer, the disturbance data corresponding to the target layer is determined based on the target category and by performing:

determining a target function by using the data input into the target layer as an independent variable, wherein the target function is positively correlated with a prediction probability corresponding to the target category;

determining target data input into the target layer in the initial prediction;

calculating a gradient value of the target function by using the target data as a value of the independent variable; and

determining the disturbance data corresponding to the target layer based on the gradient value;

inputting the feature information into the classification model for subsequent prediction, wherein the subsequent prediction includes adding the disturbance data corresponding to each of the one or more target layers to data to be input into each of the one or more target layers, a category from the plurality categories that corresponds to a maximum subsequent prediction probability obtained through the subsequent prediction remains the target category, and the maximum subsequent prediction probability is greater than the maximum initial prediction probability; and

outputting the target category and the maximum subsequent prediction probability to the service requester.

7. The computer readable medium according to claim 6, wherein determining the disturbance data corresponding to the target layer based on the gradient value includes:

obtaining a constant and a function, wherein the function is not monotonically decreasing;

substituting the gradient value as the value of the independent variable into the function to obtain a target result; and

calculating a product of the constant and the target result as the disturbance data.

8. The computer readable medium according to claim 7, wherein the function is a sign function.

9. The computer readable medium according to claim 6, wherein at least one of the one or more target layers is one of the plurality of intermediate layers.

10. The computer readable medium according to claim 6, wherein the service object includes at least one of a user, a merchant, a product, or an event.

11. A system, comprising:

one or more processors; and

memory storing contents that, when executed by the one or more processors, cause the system to perform actions comprising:

obtaining feature information of a service object from a service requester through a predetermined or dynamically-generated interface;

inputting the feature information of the service object into a service classification model for initial prediction to obtain respective initial prediction probability corresponding to each category of a plurality of categories, wherein the service classification model includes a deep neural network including an input layer, a plurality of intermediate layers, and an output layer that are sequentially stacked;

determining, from the plurality of categories, a category corresponding to a maximum initial prediction probability as a target category;

determining disturbance data corresponding to each of one or more target layers based on the target category, wherein the one or more target layers is selected from the input layer, the plurality of intermediate layers, and the output layer, and wherein for each target layer, the disturbance data corresponding to the target layer is determined based on the target category and by performing:

determining a target function by using the data input into the target layer as an independent variable, wherein the target function is positively correlated with a prediction probability corresponding to the target category;

determining target data input into the target layer in the initial prediction;

calculating a gradient value of the target function by using the target data as a value of the independent variable; and

determining the disturbance data corresponding to the target layer based on the gradient value;

inputting the feature information into the classification model for subsequent prediction, wherein the subsequent prediction includes adding the disturbance data corresponding to each of the one or more target layers to data to be input into each of the one or more target layers, a category from the plurality categories that corresponds to a maximum subsequent prediction probability obtained through the subsequent prediction remains the target category, and the maximum subsequent prediction probability is greater than the maximum initial prediction probability; and

outputting the target category and the maximum subsequent prediction probability to the service requester.

12. The system according to claim 11, wherein determining the disturbance data corresponding to the target layer based on the gradient value includes:

obtaining a constant and a function, wherein the function is not monotonically decreasing;

substituting the gradient value as the value of the independent variable into the function to obtain a target result; and

calculating a product of the constant and the target result as the disturbance data.

13. The system according to claim 12, wherein the function is a sign function.

14. The system according to claim 11, wherein at least one of the one or more target layers is one of the plurality of intermediate layers.

15. The system according to claim 11, wherein the service object includes at least one of a user, a merchant, a product, or an event.