

METHOD AND APPARATUS FOR PREVENTING MODEL THEFT DURING MODEL-BASED CLASSIFICATION

ABSTRACT

Embodiments of the present specification provide a method, apparatus, and electronic device for preventing model theft during classification using a service classification model. The service classification model is a deep neural network for predicting a category of a service object, including an input layer, multiple intermediate layers, and an output layer that are sequentially stacked. The method includes: obtaining feature information of a service object; inputting the feature information into the service classification model for initial prediction to obtain a category corresponding to a maximum initial prediction probability as a target category; determining, based on the target category, disturbance data corresponding to a target layer selected in advance from the input layer, the multiple intermediate layers, and the output layer; inputting the feature information into the service classification model for subsequent prediction, including adding the corresponding disturbance data to data to be input into the target layer; and outputting the target category and a maximum subsequent prediction probability. This can prevent model theft while ensuring that the prediction results are still usable.

FIG. 2