

CLAIMS

~~What is claimed is:~~

1. A ~~classification model generation~~computer implemented method, for identifying money laundering risks, the method comprising:

5 obtaining (602) a ~~training sample to be identified~~ dataset, wherein the ~~training sample to be identified~~ dataset comprises at least two pieces of ~~sample to be identified~~ data ~~and a sample label corresponding to each piece of sample data;~~

constructing (604) a first relational network of the at least two pieces of ~~sample to be identified~~ data, and representing the at least two pieces of ~~sample to be identified~~ data by using nodes of the first relational network;

10 wherein each node in the first relational network is connected to a set of adjacent nodes through a set of outgoing edges, each outgoing edge comprising a weight value, wherein the weight value represents an amount of money transferred between accounts;

15 obtaining (606) a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model; ~~and, comprising:~~ training for each node in the first relational network,

calculating a ~~classification model by using the training sample dataset to obtain~~ transition probability for each outgoing edge of the node based on i) the weight value of the outgoing edge and ii) a sum of weight values of the set of outgoing edges;

20 generating a first node sequence using a random walk process based on the transition probability of each outgoing edge; and

vectorizing each node in the first node sequence based on the predetermined network embedding learning model; and

25 determining (608), based on a pre-trained classification model, ~~wherein the classification model associates the sample data~~ label ~~with~~ corresponding to the vectorized representation of each node in the first relational network;

30 2. The method according to claim 1, wherein ~~training the pre-trained classification model comprises~~ obtaining a training sample dataset ~~comprises:~~

~~obtaining the training sample dataset~~ based on a predetermined time interval.

3

2. The method according to claim 1, wherein obtaining (606) a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model comprises:

- 5 performing sequence sampling on each node in the first relational network by using a random walk algorithm, and generating a first node sequence; and
vectorizing each node in the first node sequence based on the predetermined network embedding learning model.

4

- 10 3. The method according to claim 32, wherein vectorizing each node in the first node sequence based on the predetermined network embedding learning model comprises:

vectorizing (708) each node in the first node sequence based on a Node2vec network embedding learning model.

15

4. The method according to claim 2, wherein vectorizing each node in the first node sequence based on the predetermined network embedding learning model comprises:

- vectorizing each node in the first node sequence based on a DeepWalk network embedding learning model.
- 20

5. The method according to claim 3, wherein vectorizing (708) each node in the first node sequence based on a Node2vec network embedding learning model comprises:

- vectorizing, by the Node2vec network embedding learning model, each node in the first node sequence based on the SkipGram framework of Word2vec.
- 25

~~each node in the first node sequence based on the predetermined network embedding learning model comprises:~~

~~vectorizing each node in the first node sequence based on a DeepWalk network embedding learning model.~~

- 30 ~~6. The method according to claim 4, wherein vectorizing each node in the first node sequence based on a Node2vec network embedding learning model comprises:~~

~~vectorizing, by the Node2vec network embedding learning model, each node in the first node sequence based on the SkipGram framework of Word2vec.~~

7

6. The method according to claim 1, wherein the ~~classification model~~ data label comprises ~~an XGBoost model, a random forest model, white data label and a support vector machine model, or a logistic regression model.~~

8. ~~A~~black data ~~identification~~label, the method, preferably further comprising:

5 ~~obtaining a to-be-identified dataset~~if the data label is a black data label, performing subsequent processing on the vectorized representation, corresponding to the data label, of the node in the first relational network.

7. The method according to claim 1, wherein the to-be-identified dataset comprises
10 ~~a set of to-be-identified accounts that determines, by using the classification model, whether there is a money laundering behavior.~~

8. The method according to one of the preceding claims, wherein the training of the pre-trained classification model comprises:

15 ~~obtaining (202) the training sample dataset, wherein the training sample dataset comprises at least two pieces of to-be-identified sample data and a sample label corresponding to each piece of sample data;~~

~~constructing (204) a second relational network of the at least two pieces of to-be-identified sample data, and representing the at least two pieces of to-be-identified sample data by using nodes of the second relational network;~~
20

~~obtaining (206) a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model; and~~

~~determining, based on training (208) a pre-trained classification model, a data by using the training sample dataset to obtain the classification model, wherein the classification model associates the sample label corresponding to with the vectorized representation of each node in the second relational network.~~
25

9. The method according to claim 8, wherein obtaining (206) a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model comprises:
30

~~performing (302) sequence sampling on each node in the second relational network by using a random walk algorithm, and generating a second node sequence; and~~

~~vectorizing (304) each node in the second node sequence based on the~~

predetermined network embedding learning model.

10. The method according to claim 9, wherein vectorizing (304) each node in the second node sequence based on the predetermined network embedding learning model
5 comprises:

vectorizing (510) each node in the second node sequence based on a Node2vec network embedding learning model.

11. The method according to claim 9, wherein vectorizing (304) each node in the
10 second node sequence based on the predetermined network embedding learning model comprises:

vectorizing each node in the second node sequence based on a DeepWalk network embedding learning model.

12. The method according to claim 10, wherein vectorizing (510) each node in the
15 second node sequence based on a Node2vec network embedding learning model comprises:

vectorizing, by the Node2vec network embedding learning model, each node in the second node sequence based on the SkipGram framework of Word2vec.

20

13. The method according to claim 8, wherein the ~~data label comprises a white data label and a black data label.~~

~~14. The method according to claim 13, further comprising:
if the data label is a black data label, performing subsequent processing on the
25 vectorized representation, corresponding to the data label, of the node in the second relational network.~~

~~15. The method according to claim 8, wherein the to-be-identified dataset comprises a set of to-be-identified accounts that determines, by using the classification model, whether there is a money laundering behavior.~~

30 ~~16. A classification comprises an XGBoost model generation apparatus, comprising:~~

~~a first acquisition module, configured to obtain a training sample dataset, wherein the training sample dataset comprises at least two pieces of sample data and a sample label corresponding to each piece of sample data;~~

~~a first construction module, configured to construct a first relational network of the at least two pieces of sample data, and represent the at least two pieces of sample data by using nodes of the first relational network;~~

~~a first learning module, configured to obtain a vectorized representation of each node in the first relational network based on a predetermined network embedding learning, a random forest model; and~~

~~a training module, configured to train a classification, a support vector machine model by using the training sample dataset to obtain the classification, or a logistic regression model, wherein the classification model associates the sample label with the vectorized representation of each node in the first relational network.~~

~~17. The apparatus according to claim 16, wherein the first learning module comprises:~~

~~a first generation submodule, configured to perform sequence sampling on each node in the first relational network by using a random walk algorithm, and generate a first node sequence; and~~

~~a second learning submodule, configured to vectorize each node in the first node sequence based on the predetermined network embedding learning model.~~

~~18. A data identification apparatus, comprising:~~

~~a second acquisition module, configured to obtain a to-be-identified dataset, wherein the to-be-identified dataset comprises at least two pieces of to-be-identified data;~~

~~a second construction module, configured to construct a second relational network of the at least two pieces of to-be-identified data, and represent the at least two pieces of to-be-identified data by using nodes of the second relational network;~~

~~a third learning module, configured to obtain a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model; and~~

~~an identification module, configured to determine, based on a pre-trained classification model, a data label corresponding to the vectorized representation of each node in the second relational network.~~

~~19. The apparatus according to claim 18, wherein the third learning module comprises:~~

~~a second generation submodule, configured to perform sequence sampling on each node in the second relational network by using a random walk algorithm, and generate~~

~~a second node sequence; and~~

~~a fourth learning submodule, configured to vectorize each node in the second node sequence based on the predetermined network embedding learning model.~~

~~20~~

- 5 14. A computing device, comprising a memory, a processor, and computer instructions that are stored in the memory and that can run on the processor, wherein steps of the method according to any one of the preceding claims ~~1 to 7 or 8 to 15~~ are implemented when the processor executes the instructions.

- 10 ~~21. A computer readable storage medium on which computer instructions are stored, wherein steps of the method according to any one of claims 1 to 7 or 8 to 15 are implemented when the computer instructions are executed by a processor.~~