CLASSIFICATION MODEL GENERATION METHOD AND APPARATUS, AND DATA IDENTIFICATION METHOD AND APPARATUS

TECHNICAL FIELD

[0001] The present application relates to the field of computer data security technologies, and in particular, to classification model generation methods and apparatuses, data identification methods and apparatuses, computing devices, and computer storage media.

BACKGROUND

[0002] Currently, the anti-money laundering industry identifies suspicious transactions such as crimes by manually designing service features to construct a rule model. A conventional relational network data (for example, a fund network or a same device network) feature is normally obtained through manual construction. For example, statistics on network nodes or edges can be constructed to characterize subgraph patterns of nodes. This type of feature has no good effect on node classification. In addition, this type of statistics-based feature can only describe a pattern of a subgraph in a local (one hop) relationship of a node, and cannot comprehensively represent a subgraph in a multi-hop relationship and valid information may be missing.

SUMMARY

[0003] In view of this, embodiments of the present application provide classification model generation methods and apparatuses, data identification methods and apparatuses, computing devices, and computer storage media, so as to alleviate a technical defect in the existing technology.

[0004] Some embodiments of the present application disclose a classification model generation method, including: obtaining a training sample dataset, where the training sample dataset includes at least two pieces of sample data and a sample label corresponding to each piece of sample data; constructing a first relational network of the at least two pieces of sample data, and representing the at least two pieces of sample

data by using nodes of the first relational network; obtaining a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model; and training a classification model by using the training sample dataset to obtain the classification model, where the classification model associates the sample label with the vectorized representation of each node in the first relational network.

[0005] According to another aspect, some embodiments of the present application further provide a data identification method, including: obtaining a to-be-identified dataset, where the to-be-identified dataset includes at least two pieces of to-be-identified data; constructing a second relational network of the at least two pieces of to-be-identified data, and representing the at least two pieces of to-be-identified data by using nodes of the second relational network; obtaining a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model; and determining, based on a pre-trained classification model, a data label corresponding to the vectorized representation of each node in the second relational network.

[0006] According to another aspect, some embodiments of the present application further provide a classification model generation apparatus, including: a first acquisition module, configured to obtain a training sample dataset, where the training sample dataset includes 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 model; and a training module, configured to train a classification model by using the training sample dataset to obtain the classification model, where the classification model associates the sample label with the vectorized representation of each node in the first relational network.

[0007] According to another aspect, some embodiments of the present application further provide a data identification apparatus, including: a second acquisition module, configured to obtain a to-be-identified dataset, where the to-be-identified dataset includes 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.

[0008] According to another aspect, the present application further provides a computing device, including a memory, a processor, and computer instructions that are stored in the memory and that can run on the processor. When executing the computer instructions, the processor implements the steps of the classification model generation method or the data identification method.

[0009] According to another aspect, the present application further provides a computer readable storage medium, where the computer readable storage medium stores computer instructions, and the computer instructions are executed by a processor to implement the steps of the classification model generation method or the data identification method.

[0010] The present application provides classification model generation methods and apparatuses, and data identification methods and apparatuses, where the data identification method includes: obtaining a to-be-identified dataset; constructing a second relational network of the at least two pieces of to-be-identified data, and representing the at least two pieces of to-be-identified data by using nodes of the second relational network; obtaining a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model; and determining, based on a pre-trained classification model, a data label corresponding to the vectorized representation of each node in the second relational network.

BRIEF DESCRIPTION OF DRAWINGS

[0011] FIG. 1 is a schematic structural diagram illustrating a computing device, according to some embodiments of the present specification;

[0012] FIG. 2 is a flowchart illustrating a classification model generation method, according to some embodiments of the present specification;

[0013] FIG. 3 is a flowchart illustrating a classification model generation method, according to some embodiments of the present specification;

[0014] FIG. 4 is a flowchart illustrating a classification model generation method, according to some embodiments of the present specification;

[0015] FIG. 5 is a schematic diagram illustrating a relational network structural diagram and a vectorized representation of each node in the relational network structural diagram, according to some embodiments of the present specification;

[0016] FIG. 6 is a flowchart illustrating a data identification method, according to some embodiments of the present specification;

[0017] FIG. 7 is a flowchart illustrating a data identification method, according to some embodiments of the present specification;

[0018] FIG. 8 is a schematic structural diagram illustrating a classification model generation apparatus, according to some embodiments of the present specification; and

[0019] FIG. 9 is a schematic structural diagram illustrating a data identification apparatus, according to some embodiments of the present specification.

DESCRIPTION OF EMBODIMENTS

[0020] Many specific details are described in the following description to facilitate full understanding of the present application. However, the present application can be implemented in many different methods from those described here. A person skilled in the art can perform similar promotion without departing from the intension of the present application. Therefore, the present application is not limited to the specific embodiment disclosed below.

[0021] Terms used in some embodiments of the present specification are merely used to describe specific embodiments, and are not intended to limit the some embodiments of the present specification. The terms "a" and "the" of singular forms used in some embodiments of the present specification and the appended claims are also intended to include plural forms, unless otherwise specified in the context clearly. It should also be understood that the term "and/or" used in some embodiments of the present specification refers to including any or all possible combinations of one or more associated listed items.

[0022] It should be understood that although terms "first", "second", etc. may be used in some embodiments of the present specification to describe various types of

information, the information is not limited to these terms. These terms are only used to differentiate between information of the same type. For example, without departing from the scope of some embodiments of the present specification, first can also be referred to as second, and similarly, the second can be referred to as the first. Depending on the context, for example, the word "if" used here can be explained as "while", "when", or "in response to determining".

[0023] First, terms related to some embodiments of the present application are explained.

[0024] Anti-money laundering: Financial institutions control money laundering risks in the system through processes, rules or models.

[0025] Node2vec: A method for node vectorization in a relational network based on the Word2vec model.

[0026] In some embodiments of the present specification, classification model generation methods and apparatuses, data identification methods and apparatuses, computing devices, and computer storage media are provided, and are described in detail in the following embodiments one by one.

[0027] Referring to FIG. 1, some embodiments of the present specification provide a structural block diagram illustrating a computing device 100. Components of the computing device 100 include but are not limited to a memory 110, a processor 120, and a communications interface 140. The processor 120 is connected to the memory 110 by using a bus 130. A database 150 is configured to store a training sample dataset or a to-be-identified dataset. A network 160 is configured to generate a classification model, and is communicably coupled to the computing device 100 by using the communications interface 140.

[0028] The communications interface 140 enables the computing device 100 to communicate through one or more networks. Examples of these networks include a local area network (LAN), a wide area network (WAN), a personal area network (PAN), or a combination of communications networks such as the Internet. The network interface can include one or more of any type of wired or wireless network interface (e.g., a network interface card (NIC)), such as an IEEE 802.11 wireless local area network (WLAN) wireless interface, a global microwave interconnection access (Wi-MAX) interface, an Ethernet interface, a universal serial bus (USB) interface, a cellular network interface, a Bluetooth interface, a near field communication (NFC) interface, etc.

[0029] The memory 110 is configured to store a training sample dataset or a to-be-identified dataset that is sent by the communications interface 140 by using the bus 130, and computer instructions that are stored in the memory 110 and can run on the processor 120.

[0030] The processor 120 is configured to: after obtaining the training sample dataset or the to-be-identified dataset stored in the memory 110, execute the computer instructions stored in the memory 110, so as to generate the classification model and identify the data.

[0031] The computing device 100 can be any type of fixed or mobile computing device, including a mobile computer or mobile computing device (e.g., a tablet computer, a personal digital assistant, a laptop computer, a notebook computer, a netbook, etc.), a mobile phone (e.g., a smartphone), a wearable computing device (e.g., a smartwatch, smart glasses, etc.), or another type of mobile device, or a fixed computing device such as a desktop computer or a PC.

[0032] The processor 120 can perform the steps in the method shown in FIG. 2. FIG. 2 is a schematic flowchart illustrating a classification model generation method, according to some embodiments of the present specification, including step 202 to step 208.

[0033] Step 202: Obtain a training sample dataset, where the training sample dataset includes at least two pieces of sample data and a sample label corresponding to each piece of sample data.

[0034] In some embodiments of the present specification, the sample data includes but is not limited to white sample data and black sample data. The sample label includes but is not limited to a white sample label and a black sample label.

[0035] In practice, the training sample dataset is formed by combining the white sample data and a corresponding white sample label with the black sample data and a corresponding black sample label.

[0036] When the classification model generation method is applied to an anti-money laundering crime identification field, the black sample data can be an illegal account in which a money laundering behavior exists, and the white sample data can be a legal account in which no money laundering behavior exists; the black sample data can be obtained by using an external public information or by designing a service feature based on expert experience, and the white sample data can be obtained by excluding known black sample data from all samples.

[0037] In practice, the quantity of white sample data is far greater than the quantity of black sample data. To avoid an imbalance in sample data in the training sample dataset due to an excessive quantity of white sample data, the ratio of white sample data to black sample data is controlled to be 10:1 to 100:1. Therefore, the white sample data can be obtained by using down-sampling. Using the previous ratio between the white sample data and the black sample data can avoid overfitting of the classification model and negative effects that reduce the learning capability of the classification model.

[0038] Step 204: 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.

[0039] In some embodiments of the present specification, the first relational network includes a series of nodes and relationships, and there can be interaction between nodes, etc. Therefore, a relationship exists between nodes, and relationship construction is derived.

[0040] In some embodiments of the present specification, the at least two pieces of sample data represent nodes of the first relational network, and the relationship between the at least two pieces of sample data represent interaction between nodes.

[0041] For example, the first relational network is a static fund relational network, and the construction of the first relational network of the at least two pieces of sample data is described in detail.

[0042] For example, the at least two samples are obtained accounts that a fund transaction is performed in the last 90 days, and then fund transactions among all accounts are summarized. If the static fund relational network is a directed graph or a weighted graph, a final static fund relational network presents a ternary form, that is, U, V, and W, which indicate that there is a directional edge with a weight W from U to V, which is represented in a service scenario corresponding to the static fund relational network as follows: Account U pays W Yuan to account V. Correspondingly, if the first relational network is a same device relational network of an undirected graph or unweighted graph, in a service scenario corresponding to the same device relational network, it is indicated that both account U and account V use the same device to perform a fund transaction. Therefore, an edge from V to U does not need to be added, and W is set to 1.

[0043] Step 206: Obtain a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model.

[0044] Referring to FIG. 3, in some embodiments of the present specification, obtaining a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model includes step 302 to step 304.

[0045] Step 302: Perform sequence sampling on each node in the first relational network by using a random walk algorithm, and generating a first node sequence.

[0046] Step 304: Vectorize each node in the first node sequence based on the predetermined network embedding learning model.

[0047] In some embodiments of the present specification, vectorizing each node in the first node sequence based on the predetermined network embedding learning model includes: vectorizing each node in the first node sequence based on a Node2vec network embedding learning model; or vectorizing each node in the first node sequence based on a DeepWalk network embedding learning model.

[0048] Vectorizing each node in the first node sequence based on a Node2vec network embedding learning model includes: vectorizing, by the Node2vec network embedding learning model, each node in the first node sequence based on the SkipGram framework of Word2vec.

[0049] In some embodiments of the present specification, based on the Node2vec network embedding learning model, each node in the constructed first relational network is converted into a node sequence after sampling by using a random walk algorithm, and then, based on a SkipGram framework in the Word2vec model, probability learning and inference are further performed on the node sequence after sampling to finally obtain a vectorized representation of each node in the first relational network. A node vectorized representation obtained by using network embedding learning can enrich a relationship between nodes and improve the processing rate and effect of the classification model.

[0050] Referring to FIG. 4, the left side is a relational network structural diagram formed by 13 nodes and edges constructed by relationships among the nodes, and the relational network structural diagram is calculated based on the network embedding learning model to obtain a vectorized representation of each of the 13 nodes in the relational network structural diagram on the right side. That is, the relational network structure can be transformed into a multi-dimensional vector through a series of processes. Therefore, complex relational network information can be converted into a structured multi-dimensional feature, and a more convenient algorithm application can be implemented by using a machine learning method.

[0051] Step 208: Train a classification model by using the training sample dataset to obtain the classification model, where the classification model associates the sample label with the vectorized representation of each node in the first relational network.

[0052] In some embodiments of the present specification, the classification model includes an XGBoost model, a random forest model, a support vector machine (SVM) model, or a logistic regression (LR) model.

[0053] In some embodiments of the present specification, the training sample dataset can be further obtained based on a predetermined time interval. In this way of periodically collecting the training sample dataset and training the classification model, a new feature can be adaptively found, and validity of the classification model is continuously ensured.

[0054] The predetermined time interval can be set based on actual needs. For example, the predetermined time interval can be set to one week, one month, or three months, which is not limited in the present application.

[0055] In some embodiments of the present specification, according to the classification model generation method, the classification model can determine the first relational network feature adaptively by collecting the data sample periodically based on the predetermined time interval, and the classification model can be trained by obtaining the vectorized representation of each node in the first relational network by using the network embedding learning model. As such, working efficiency can be improved, and a network feature pattern of each node in the first relational network can be comprehensively described based on the vectorized representation of each node in the first relational network. In addition, the problem that the training sample dataset becomes invalid can be avoided by collecting data samples periodically.

[0056] Referring to FIG. 5, some embodiments of the present specification provide a schematic flowchart illustrating a classification model generation method, including step 502 to step 514.

[0057] Step 502: Obtain black sample data and a corresponding black sample label.

[0058] Step 504: Obtain white sample data and a corresponding white sample label.

[0059] Step 506: Combine the black sample data and the corresponding black sample label with the white sample data and the corresponding white sample label to form a training sample dataset.

[0060] Step 508: Construct a relational network of the black sample data and the white sample data, and represent the black sample data and the white sample data by using nodes of the relational network.

[0061] Step 510: Obtain a vectorized representation of each node in the relational network based on a Node2vec network embedding learning model.

[0062] Step 512: Train a classification model based on the training sample dataset.

[0063] Step 514: Obtain the classification model.

[0064] In some embodiments of the present specification, according to the classification model generation method, only the black and white data samples need to be collected, then a relational network feature is constructed by using the black and white sample data, and the vectorized representation of each node in the relational network is obtained by using a network embedding learning model, so as to achieve the objective of training the classification model. As such, working efficiency can be improved, and a network feature pattern of each node in the relational network can be comprehensively described based on the vectorized representation of each node in the relational network.

[0065] Referring to FIG. 6, some embodiments of the present specification provide a schematic flowchart illustrating a data identification method, including step 602 to step 608.

[0066] Step 602: Obtain a to-be-identified dataset, where the to-be-identified dataset includes at least two pieces of to-be-identified data.

[0067] In some embodiments of the present specification, the to-be-identified dataset includes a set of to-be-identified accounts that determines, by using the classification model, whether there is a money laundering behavior.

[0068] Step 604: 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.

[0069] In some embodiments of the present specification, step 604 is the same as an operation method of step 204 in the previous embodiment, and details are omitted here for simplicity.

[0070] Step 606: Obtain a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model.

[0071] In some embodiments of the present specification, obtaining a vectorized representation of each node in the second relational network based on a predetermined network embedding learning model includes: performing sequence sampling on each node in the second relational network by using a random walk algorithm, and generating a second node sequence; and vectorizing each node in the second node sequence based on the predetermined network embedding learning model.

[0072] In some embodiments of the present specification, vectorizing each node in the second node sequence based on the predetermined network embedding learning model includes: vectorizing each node in the second node sequence based on a Node2vec network embedding learning model; or vectorizing each node in the second node sequence based on a DeepWalk network embedding learning model.

[0073] Vectorizing each node in the second node sequence based on a Node2vec network embedding learning model includes: vectorizing, by the Node2vec network embedding learning model, each node in the second node sequence based on the SkipGram framework of Word2vec.

[0074] In some embodiments of the present specification, the second relational network is used as an input, and learning is performed by using the predetermined network embedding learning model, to obtain the vectorized representation of each node in the second relational network.

[0075] For example, the predetermined network embedding learning model includes the Node2vec network embedding learning model. Obtaining the vectorized representation of each node in the second relational network by using the network embedding learning model is described in detail.

[0076] Step 1: Calculate a transition probability of each edge in the second relational network based on the second relational network structure, to obtain a transition probability graph of the second relational network.

[0077] For example, the second relational network has a node U, a node V, and a node K, where the sum of weights of outgoing edges (directed graphs) is Z, a transition probability of each outgoing edge is $P(V) = W(U, V)/Z$, and $W(U, V)$ represents an edge weight from the node U to the node V. The transition probability graph of the second relational network can be obtained in the previous way.

[0078] Step 2: Randomly walk, based on the transition probability graph of the second relational network, to generate a node sequence of each node in the second relational network.

[0079] In some embodiments of the present specification, the node sequence of each node in the second relational network constructed by random walk should satisfy the following constraint condition: Assume the current node is V, the last random walk node is U, and an adjacent node of V is K, if the shortest path distance between K and U is 0, a transition probability is W(U, V)/Z/p, where p is a model parameter; if the shortest path distance between K and U is 1, the transition probability is W(U, V)/Z; if the shortest path distance between K and U is 2, the transition probability is W(U, V)/Z/q, where q is a model parameter. The previous random transition process is repeated, and the previous constraint condition is satisfied until a sequence length reaches a specified parameter MAX_LEN, where MAX_LEN is a model parameter.

[0080] Step 3: The Node2vec network embedding learning model vectorizes each node in the node sequence based on the SkipGram framework of Word2vec.

[0081] In some embodiments of the present specification, the SkipGram framework of Word2vec is used to perform optimization and learning of the model by using stochastic gradient descent, to finally obtain the vectorized representation of each node in the second relational network.

[0082] Step 608: Determine, based on a pre-trained classification model, a data label corresponding to the vectorized representation of each node in the second relational network.

[0083] In some embodiments of the present specification, the data label includes a white data label and a black data label.

[0084] In practice, the white data label is corresponding to a legal account in which no money laundering behavior exists, and the black sample data label is corresponding to an illegal account in which money laundering behavior exists. If the vectorized representation of each node in the second relational network corresponds to a white sample data label, to-be-identified data corresponding to the vectorized representation of the node is a legal account; or if the vectorized representation of each node in the second relational network corresponds to a black sample data label, to-be-identified data corresponding to the vectorized representation of the node is an illegal account.

[0085] In some embodiments of the present specification, if the data label is a black data label, subsequent processing is performed on the vectorized representation, corresponding to the data label, of the node in the second relational network.

[0086] The subsequent processing includes but is not limited to performing account fund flow tracing or a detailed identity query and login information query of a real user corresponding to an account, which is not limited in the present application.

[0087] In some embodiments of the present specification, according to the data identification method, the local subgraph pattern of the second relational network can be comprehensively described by constructing the second relational network by using the to-be-identified dataset, then the vectorized representation of each node in the second relational network is obtained based on the predetermined network embedding learning model, and the data label corresponding to the vectorized representation of each node in the second relational network is quickly determined by using the pre-trained classification model, so as to implement fast identification of the to-be-identified data.

[0088] Referring to FIG. 7, some embodiments of the present specification provide a schematic flowchart illustrating a data identification method, including step 702 to step 710.

[0089] Step 702: Obtain at least two to-be-identified accounts.

[0090] Step 704: Use the two to-be-identified accounts to form a to-be-identified account set.

[0091] Step 706: Construct a relational network of the to-be-identified account set, and represent the to-be-identified account set by using nodes of the relational network.

[0092] Step 708: Obtain a vectorized representation of each node in the relational network based on a Node2vec network embedding learning model.

[0093] Step 710: Determine, based on a pre-trained classification model, an account label corresponding to the vectorized representation of each node in the relational network.

[0094] In some embodiments of the present specification, the method is applied to the anti-money laundering field, original information such as the relational network is used as an input, and a network embedding learning model and a pre-trained classification model are used to implement rapid identification of an account involving a criminal action.

[0095] Referring to FIG. 8, some embodiments of the present specification provide a classification model generation apparatus, including: a first acquisition module 802, configured to obtain a training sample dataset, where the training sample dataset includes at least two pieces of sample data and a sample label corresponding to each

piece of sample data; a first construction module 804, 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 806, configured to obtain a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model; and a training module 808, configured to train a classification model by using the training sample dataset to obtain the classification model, where the classification model associates the sample label with the vectorized representation of each node in the first relational network.

[0096] Optionally, the first learning module 806 includes: 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.

[0097] In some embodiments of the present specification, according to the classification model generation apparatus, the classification model can determine the first relational network feature adaptively by collecting the data sample periodically based on the predetermined time interval, and the classification model can be trained by obtaining the vectorized representation of each node in the first relational network by using the network embedding learning model. As such, working efficiency can be improved, and a network feature pattern of each node in the first relational network can be comprehensively described based on the vectorized representation of each node in the first relational network. In addition, the problem that the training sample dataset becomes invalid can be avoided by collecting data samples periodically.

[0098] Referring to FIG. 9, some embodiments of the present specification provide a data identification apparatus, including: a second acquisition module 902, configured to obtain a to-be-identified dataset, where the to-be-identified dataset includes at least two pieces of to-be-identified data; a second construction module 904, 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 906, 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 908, 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.

[0099] Optionally, the third learning module 906 includes: 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.

[0100] In some embodiments of the present specification, according to the data identification apparatus, the local subgraph pattern of the second relational network can be comprehensively described by constructing the second relational network by using the to-be-identified dataset, then the vectorized representation of each node in the second relational network is obtained based on the predetermined network embedding learning model, and the data label corresponding to the vectorized representation of each node in the second relational network is quickly determined by using the pre-trained classification model, so as to implement fast identification of the to-be-identified data.

[0101] Some embodiments of the present specification are described in a progressive method. The same or similar parts of the embodiments can be referred to each other, and each embodiment focuses on a difference from another embodiment. Particularly, a device embodiment is similar to a method embodiment, and therefore is described briefly. For a related part, references can be made to some descriptions in the method embodiment.

[0102] Some embodiments of the present application further provides a computer readable storage medium, where the computer readable storage medium stores computer instructions, and the instructions are executed by a processor to implement the steps of the classification model generation method.

[0103] The previous description is a schematic solution of a computer readable storage medium in this embodiment. It is worthwhile to note that the technical solution of the storage medium and the technical solution of the previous classification model generation method belong to the same idea. For details not described in the technical solution of the storage medium, refer to descriptions of the technical solution of the previous classification model generation method.

[0104] Some embodiments of the present application further provides a computer readable storage medium, where the computer readable storage medium stores

computer instructions, and the instructions are executed by a processor to implement the steps of the data identification method.

[0105] The previous is a schematic solution of a computer readable storage medium in this embodiment. It is worthwhile to note that the technical solution of the storage medium and the technical solution of the previous data identification method belong to the same concept. For details not described in the technical solution of the storage medium, refer to descriptions of the technical solution of the previous data identification method.

[0106] Specific embodiments of the present specification are described above. Other embodiments fall within the scope of the appended claims. In some situations, the actions or steps described in the claims can be performed in an order different from the order in the embodiments and the desired results can still be achieved. In addition, the process depicted in the accompanying drawings does not necessarily need a particular execution order to achieve the desired results. In some embodiments, multi-tasking and concurrent processing is feasible or may be advantageous.

[0107] Technical carriers related to payment in the embodiments of the present application can include, for example, Near Field Communication (NFC), Wireless Fidelity (Wi-Fi), 3G/4G/5G, a technology of swiping card by using a point of sale (POS) machine, a quick response code scanning technology, a bar code scanning technology, Bluetooth, infrared, a short message service (SMS), and a multimedia messaging service (MMS).

[0108] The computer instruction includes computer instruction code, and the computer instruction code can be in a source code form, an object code form, an executable file, some intermediate forms, etc. The computer readable medium can include any entity or apparatus that can carry the computer instruction code, a recording medium, a USB flash drive, a removable hard disk, a magnetic disk, an optical disc, a computer memory, a read-only memory (ROM), a random access memory (RAM), an electrical carrier signal, a telecommunications signal, a software distribution medium, etc. It is worthwhile to note that content included in the computer readable medium can be appropriately increased or decreased according to a need of legislation and patent practice in a jurisdiction, for example, in some jurisdictions, according to legislation and patent practice, the computer readable medium does not include an electrical carrier signal and a telecommunications signal.

[0109] It is worthwhile to note that, for ease of description, the previous method

embodiments are described as a series of action combinations. However, a person skilled in the art should understand that the present application is not limited to the described action sequence, because according to the present application, some steps can be performed in another order or at the same time. In addition, a person skilled in the art should also understand that the embodiments described in the present specification are all preferred embodiments, and involved actions and modules are not necessarily required by the present application.

[0110] In the previous embodiments, the description of each embodiment has respective focuses. For a part not described in detail in some embodiments, references can be made to related descriptions in other embodiments.

[0111] The previous disclosed preferred embodiments of the present application are merely intended to help describe the present application. The optional embodiment does not describe all details, and the present application is not merely limited to the specific embodiment method. Clearly, many modifications and changes can be made according to the content of the present specification. The present specification selects and specifically describes these embodiments to better explain principles and practical applications of the present application, so a person skilled in the art can better understand and use the present application. The present application is only limited by the claims and all the scope and equivalents thereof.

CLAIMS

What is claimed is:

1. A classification model generation method, comprising:

obtaining 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;

constructing a first relational network of the at least two pieces of sample data, and representing the at least two pieces of sample data by using nodes of the first relational network;

obtaining a vectorized representation of each node in the first relational network based on a predetermined network embedding learning model; and

training a classification model by using the training sample dataset to obtain the classification model, wherein the classification model associates the sample label with the vectorized representation of each node in the first relational network.

2. The method according to claim 1, wherein obtaining a training sample dataset comprises:

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

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

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. The method according to claim 3, 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 Node2vec network embedding learning model.

5. The method according to claim 3, 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.

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

8. A data identification method, comprising:

obtaining a to-be-identified dataset, wherein the to-be-identified dataset comprises at least two pieces of to-be-identified data;

constructing a second relational network of the at least two pieces of to-be-identified data, and representing the at least two pieces of to-be-identified data by using nodes of the second relational network;

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

determining, based on a pre-trained classification model, a data label corresponding to the vectorized representation of each node in the second relational network.

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

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

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

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

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

11. The method according to claim 9, wherein vectorizing each node in the 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 each node in the 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.

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 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.

16. A classification 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 model; and

a training module, configured to train a classification model by using the training sample dataset to obtain the classification 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. 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 claims 1 to 7 or 8 to 15 are implemented when the processor executes the instructions.

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.