

DEEP LEARNING-BASED AUTOMATIC PROCESSING METHOD AND SYSTEM FOR UNSTRUCTURED DATA

TECHNICAL FIELD

5 [0001] The present invention relates to a data processing technology, and specifically relates to a deep learning-based automatic processing method and system for unstructured data.

BACKGROUND

10 [0002] With the rapid development of information technology, the generation and storage capacity of unstructured data has increased exponentially. Unstructured data in various forms, including text, images, audio, and video, has been widely applied across various industries. However, it lacks of a fixed storage structure for the unstructured data with diverse formats and complex content, making traditional data processing methods
15 inadequate to meet the demands for efficient classification, labelling, and utilization of such data. Especially in scenarios with massive data volumes and frequent updates, there are significant challenges in quickly realizing the exploration and utilization of data value.

 [0003] Existing technologies mostly rely on manual or semi-automated methods to process the unstructured data, but these methods are often inefficient and costly, and are
20 difficult to realize automatic and precise processing especially in stages of data quality assessment and deep feature extraction. Moreover, there is still a lack of effective model architectures for comprehensive exploration of both implicit features (such as association rules and change trends) and recessive features (such as time series and numerical distribution). Furthermore, during the process of constructing inference rules based on
25 feature association relationships, current technologies cannot fully utilize domain knowledge and causal inference methods, resulting in limited effectiveness in data optimization and structured transformation.

 [0004] To solve the above problems, there is an urgent need for a deep learning-based automatic processing method for unstructured data, which can automate the entire process
30 from data collection, classification and labelling to deep feature mining, rule inferring, and optimization processing, thereby improving the efficiency and quality of data processing.

SUMMARY OF THE INVENTION

 [0005] The embodiments of the present invention provide a deep learning-based

automatic processing method and system for unstructured data, which can solve the problems in related technology.

[0006] In a first aspect,

[0007] the deep learning-based automatic processing method for unstructured data
5 provided by the embodiment of the present invention includes:

[0008] acquiring unstructured data by an intelligent data collector; constructing a data
feature fingerprint library; performing classification and labelling on the unstructured data
based on the data feature fingerprint library to obtain classified and labelled data;
constructing a data quality assessment model by an incremental learning algorithm, the data
10 quality assessment model scoring the classified and labelled data from three dimensions of
data integrity, timeliness and value density to obtain a scoring result; constructing a priority
processing queue according to the scoring result, and performing data cleaning and
standardized processing on the classified and labelled data in the priority processing queue
to obtain a standardized training data set;

15 [0009] performing deep learning processing on the standardized training data set by a
multi-level cascade network structure, the multi-level cascade network structure identifying
target features in the standardized training data set through a self-attention mechanism
according to class information of the classified and labelled data; constructing a feature
association graph based on the target features; constructing a bidirectional inference
20 channel on the feature association graph, the bidirectional inference channel including a
forward channel and a reverse channel, the forward channel extracting numerical
distribution features, time-order features and attribute features in the standardized training
data set by a convolutional neural network so as to obtain dominant feature data, and the
reverse channel extracting data item association rules, data change trends and data
25 interaction modes in the standardized training data set through a recurrent neural network
so as to obtain recessive feature data; performing feature combination on the dominant
feature data and the recessive feature data by a dynamic weight allocation strategy to obtain
a multi-dimensional feature tensor; and

[0010] performing rule matching on the multi-dimensional feature tensor based on a
30 preset domain knowledge base to obtain an inference rule set; performing hierarchical
processing on the multi-dimensional feature tensor by the inference rule set so as to obtain
initial structured data; analyzing the initial structured data by a causal inference method to
obtain a decision basis chain; generating optimization parameters based on the decision
basis chain; and performing optimization processing on the initial structured data according
35 to the optimization parameters and the class information of the classified and labelled data

to obtain final structured data.

[0011] In an optional embodiment,

[0012] the acquiring unstructured data by an intelligent data collector; constructing a data feature fingerprint library; performing classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data includes:

[0013] acquiring the unstructured data by the intelligent data collector, the intelligent data collector including a local feature channel and a global feature channel, the local feature channel performing word-level feature extraction and sentence-level feature extraction by a hierarchical attention network to obtain local association features, the global feature channel constructing a data relationship graph by a graph neural network and extracting global relationship features by message transfer, and performing attention fusion on the local association features and the global relationship features to obtain fused feature data;

[0014] performing multi-head attention computation on the fused feature data to obtain feature importance weights, screening the fused feature data according to the feature importance weights, mapping the screened features by a dynamic locality sensitive hashing algorithm to generate feature fingerprints, and constructing the feature fingerprints into the data feature fingerprint library that supports nearest neighbor query;

[0015] performing data enhancement on the fused feature data to obtain an enhanced feature set; computing contrast loss of sample pairs based on the enhanced feature set to obtain a semantic similarity matrix; constructing a data association graph through the nearest neighbor query result of the data feature fingerprint library and the semantic similarity matrix; transmitting known labels through graph attention on the data association graph to obtain an initial classification and labelling result, and extracting an entity relationship according to the data association graph to construct a domain knowledge graph;

[0016] constructing a meta-learner, performing assessment modeling on a classification result of each sample by taking a feature vector and confidence of the initial classification and labelling result as inputs to obtain a classification reliability score, tagging samples which are less than a preset reliability threshold and verified by the data feature fingerprint library as first-class to-be-re-checked data, performing logic verification on the initial classification and labelling result by the domain knowledge graph, and tagging samples which do not pass the verification as second-class to-be-re-checked data;

[0017] modeling a selection process of the first-class to-be-re-checked data and the

second-class to-be-re-checked data as a Markov decision process, constructing a hybrid reward function according to the classification reliability score and a sample information gain, optimizing a sample selection strategy by deep reinforcement learning, and selecting an optimal to-be-re-checked sample for manual labelling to obtain a manual labelling result;

5 and

[0018] performing weighted fusion on the manual labelling result and the initial classification and labelling result, performing consistency verification on the labelling result subjected to the weighted fusion by a similarity index of the data feature fingerprint library, and generating classified and labelled data containing the fusion feature data, the semantic similarity matrix, the classification reliability score and the domain knowledge graph verification result.

10

[0019] In an optional embodiment,

[0020] the constructing a data quality assessment model by an incremental learning algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; constructing a priority processing queue according to the scoring result, and performing data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set includes:

15

[0021] acquiring timestamp information of the classified and labelled data, performing computation by an exponential decay function to obtain a data time attenuation coefficient, and performing weighting processing on the classified and labelled data based on the data time attenuation coefficient to obtain a sample weight;

20

[0022] constructing the data quality assessment model, constructing a loss function by taking the sample weight as a weight item, optimizing the loss function by an online stochastic gradient descent algorithm until the model is converged, so as to obtain a trained data quality assessment model, and performing quality assessment on the classified and labelled data by the trained data quality assessment model;

25

[0023] computing an effective feature ratio based on the quality assessment result to obtain a data integrity score, obtaining a data timeliness score according to the data time attenuation coefficient, performing computation in combination with the sample weight to obtain a data value density score, and performing weighted fusion on the data integrity score, the data timeliness score and the data value density score to obtain a data quality score;

30

[0024] constructing the priority processing queue according to the data quality score, tagging the classified and labelled data with the data quality score less than a first preset

35

threshold as to-be-removed data, tagging the classified and labelled data with the data quality score greater than a second preset threshold as priority processing data, computing the data processing priority based on the data quality score and the sample weight, and generating the priority data queue with processing tags;

5 [0025] according to the priority sequence of the priority data queue, carrying out unified transformation of field names and data formats on the classified and labelled data to obtain standardized features, constructing a time-order completion model based on the data timeliness score and the data integrity score, and completing missing fields by the time-order completion model to obtain structured features;

10 [0026] constructing a feature fingerprint according to the structured feature and the data value density score, identifying duplicated data based on the feature fingerprint, carrying out anomaly detection and de-noising processing in combination with the data quality score, and generating a quality processing record and cleaned feature data; and

 [0027] carrying out standardized transformation on numerical features, class features
15 and text features in the cleaned feature data respectively, guiding word vector mapping of the text features by the feature fingerprint, and combining the transformed standardized features with the data quality score, the data processing priority and the quality processing record to generate the standardized training data set.

 [0028] In an optional embodiment,

20 [0029] the performing deep learning processing on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and labelled data includes:

 [0030] constructing the multi-level cascade network structure, the multi-level cascade
25 network structure including a feature extraction level, an attention level and a feature fusion level, the feature extraction level consisting of a plurality of convolutional blocks, the attention level adopting a self-attention mechanism, and the feature fusion level being used for feature dimension reduction and combination;

 [0031] performing class information extraction on the classified and labelled data in
30 the standardized training data set, constructing a class embedding matrix, and performing computation based on the class embedding matrix and the standardized training data set to obtain a class mapping matrix;

 [0032] processing the standardized training data set by the feature extraction level of
the multi-level cascade network structure, performing product operation on the class
35 mapping matrix and extracted features to obtain class perception features, and splicing the

class perception features extracted at different levels in a channel dimension to obtain multi-scale cascade features;

[0033] in the attention level, generating a query matrix, a key matrix and a value matrix based on the multi-scale cascade features and the class embedding matrix,
5 computing the similarity of the query matrix and the key matrix through the self-attention mechanism to obtain an attention weight, and multiplying the attention weight by the value matrix to obtain attention features;

[0034] in the feature fusion level, performing residual connection and normalization processing on the attention features to obtain standardized features, constructing a gating
10 weight based on the standardized features and the class embedding matrix, and performing weighting on the gating weight and the standardized features to obtain enhanced features; and

[0035] transmitting the enhanced features of each level in a level-by-level way in the multi-level cascade network structure and performing feature cascade, and performing
15 cross-layer feature aggregation to obtain the target features in the standardized training data set.

[0036] In an optional embodiment,

[0037] the performing rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set includes:

[0038] performing dimension reduction processing on the multi-dimensional feature
20 tensor by a feature dimension reduction network to obtain a feature sequence, constructing a feature index tree on the feature sequence, setting a dynamically sized sliding window based on the feature index tree, adaptively adjusting the size of the sliding window according to the feature sequence, and performing sliding scanning on the feature sequence
25 to extract a feature mode;

[0039] inputting the feature mode into a pre-training text encoder to obtain a feature encoding vector, and inputting a rule mode in the preset domain knowledge base into the pre-training text encoder to obtain a rule encoding vector; computing cosine similarity between the feature encoding vector and the rule encoding vector as a matching degree;

[0040] computing a mean value and a standard deviation of the matching degree,
30 determining an adaptive threshold coefficient according to a preset confidence interval, performing weighting on the mean value and the standard deviation based on the adaptive threshold coefficient to obtain an adaptive threshold, screening the rule modes with the matching degrees greater than the adaptive threshold, and forming a candidate rule set by
35 the screened rule modes and corresponding structured templates; and

[0041] performing dependency analysis on rules in the candidate rule set, establishing a dependency relationship among the rules based on input features and output features of the rules, representing the dependency relationship as a directed graph, nodes representing the rules and directed edges representing the dependency relationship, computing an in-degree and an out-degree of each rule node, determining a dependency level of the rules, and allocating the rules having the direct dependency relationship to adjacent levels by an improved topological sorting algorithm, so as to obtain the inference rule set.

[0042] In an optional embodiment,

[0043] the performing hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data includes:

[0044] extracting the feature vectors of rule-related regions from the feature sequence for the rules at each rule level in the inference rule set, inputting the feature vectors into a rule matching network, the rule matching network including a feature transformation layer and a matching degree computation layer, outputting rule applicability scores, and standardizing the rule applicability scores to serve as rule confidence;

[0045] performing descending on the rules at each rule level based on the rule confidence to obtain a sorting rule set, applying the rules to the corresponding structured templates in sequence according to the sorting sequence, and generating structured fragments by feature mapping, the structured templates defining a mapping relationship between the features and the structured data;

[0046] detecting action regions of the rules in the sorting rule set, when the action regions of a plurality of rules are overlapped, extracting a range of an overlapped region, computing a coverage degree of each rule in the overlapped region, taking a weighted sum of the rule confidence and the coverage degrees as a rule priority, and selecting the rule with the highest priority to be applied to the overlapped region, thereby ensuring to generate a unique structured fragment;

[0047] constructing a multi-level verification network containing an attribute verification level, a relationship verification level and a value domain verification level, verifying the integrity of the structured fragment based on an attribute graph in the attribute verification level, verifying the relationship reasonability between the structured fragments based on a knowledge graph in the relationship verification level, verifying the validity of attribute values in the structured fragments based on a statistical model in the value domain verification level, and integrating verification result of each level to obtain verified structured fragments; and

[0048] extracting semantic features in the verified structured fragments, computing the

semantic similarity among the fragments, constructing hierarchical organization structures of the fragments based on the semantic similarity, and integrating the structured fragments related to semantics according to the hierarchical organization structures to generate the initial structured data meeting a preset specification.

5 [0049] In an optional embodiment,

[0050] the analyzing the initial structured data by a causal inference method to obtain a decision basis chain; generating optimization parameters based on the decision basis chain; and performing optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to
10 obtain final structured data includes:

[0051] computing a partial association coefficient for a variable pair in the initial structured data, performing a conditional independence test based on the partial association coefficient to obtain a conditional independence relationship among variables, and constructing a causality graph according to the conditional independence relationship;

15 [0052] constructing a linear structure equation model based on the causality graph, optimizing parameters of the linear structure equation model by maximum likelihood estimation to obtain a structural causal model, the structural causal model being used for quantifying causal effect intensity among the variables in the causality graph;

[0053] computing the number of associated edges of each node in the causality graph,
20 determining the nodes with the number of the associated edges exceeding a preset associated edge number threshold as decision nodes, tracing associated nodes level by level from the decision nodes along causality edges in the causality graph to obtain factor chains, computing a cumulative causal effect of the factor chains based on the structural causal model, and determining the factor chains with the cumulative causal effect exceeding a
25 preset effect threshold as candidate decision basis chains;

[0054] in the structural causal model, changing the value of each node in the candidate decision basis chains and substituting into the structural causal model to obtain a corresponding downstream node predicted value after the value of the node is changed, determining the stability of the candidate decision basis chains based on a change
30 amplitude of the predicted value, and determining the candidate decision basis chains with the stability meeting a preset condition as decision basis chains;

[0055] computing an influence degree of each node in the decision basis chains on a target variable based on the structural causal model, screening nodes with the influence degree exceeding a preset influence threshold as a feature set according to the influence
35 degree, taking the influence degree of each node in the feature set as a feature importance

score, and combining the feature importance score with the class discrimination of the classified and labelled data to obtain a feature optimization parameter;

[0056] performing weighted adjustment on the value of the corresponding feature in the initial structured data based on the feature optimization parameter to obtain optimized data, the feature optimization parameter being used for determining a weight coefficient of feature adjustment; and

[0057] computing the feature distribution difference between the optimized data and the initial structured data to obtain an optimization effect index; inputting the optimized data into the structural causal model to verify causal relationship constraint so as to obtain a verification result, performing iterative optimization on the optimized data according to the optimization effect index and the verification result, and when the optimization index reaches a preset target and the verification result meets causal consistency, determining the current optimized data as the final structured data.

[0058] In a second aspect,

[0059] the deep learning-based automatic processing system for unstructured data provided by the embodiment of the present invention includes:

[0060] a first unit, configured to acquire unstructured data by an intelligent data collector; construct a data feature fingerprint library; perform classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data; construct a data quality assessment model by an incremental learning algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; construct a priority processing queue according to the scoring result, and perform data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set;

[0061] a second unit, configured to perform deep learning processing on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and labelled data; construct a feature association graph based on the target features; construct a bidirectional inference channel on the feature association graph, the bidirectional inference channel including a forward channel and a reverse channel, the forward channel extracting numerical distribution features, time-order features and attribute features in the standardized training data set by a convolutional neural network so as to obtain dominant feature data, and the reverse channel extracting data item association rules, data change

trends and data interaction modes in the standardized training data set through a recurrent neural network so as to obtain recessive feature data; and perform feature combination on the dominant feature data and the recessive feature data by a dynamic weight allocation strategy to obtain a multi-dimensional feature tensor; and

5 [0062] a third unit, configured to perform rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set; perform hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data; analyze the initial structured data by a causal inference method to obtain a decision basis chain; generate optimization parameters based
10 on the decision basis chain; and perform optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to obtain final structured data.

[0063] In a third aspect,

[0064] an electronic device provided by the embodiment of the present invention
15 includes:

[0065] a processor; and

[0066] a memory configured to store processor executable instructions; and

[0067] the processor is configured to call instructions stored in the memory to execute
the abovementioned method.

20 [0068] In a fourth aspect,

[0069] the embodiment of the present invention provides a computer-readable storage medium having computer program instructions stored therein, when executed by a processor, the computer program instructions execute the abovementioned method.

[0070] In the embodiments, by constructing the data feature fingerprint library and in
25 combination with the incremental learning algorithm, the data quality assessment model is established, the data is scored from three dimensions of the data integrity, timeliness and value density, and therefore, the priority processing queue is established, which can effectively improve the data processing efficiency and accuracy, thereby ensuring that important data can be preferentially processed; and moreover, the quality of training data is
30 ensured by data cleaning and normalized processing. The deep learning processing is performed by the multi-level cascade network structure in combination with the self-attention mechanism; the dominant features and the recessive features are extracted by the bidirectional inference channel, and feature combination is performed by the dynamic weight allocation strategy, therefore, various feature information in the data can be
35 comprehensively captured, the integrity and accuracy of feature extraction can be improved,

and the understanding capability of the model to the complex unstructured data is enhanced. The rule matching and causal inference are performed based on the preset domain knowledge base, and the optimization process is guided by the decision basis chains, so that intelligent transformation from the unstructured data to the structured data is realized, the interpretability and reliability of data processing are improved, the processing result better meets actual application demands, and high-quality data support is provided for subsequent data analysis and decision.

BRIEF DESCRIPTION OF THE DRAWINGS

[0071] FIG. 1 is a flowchart of a deep learning-based automatic processing method for unstructured data according to an embodiment of the present invention; and

[0072] FIG. 2 is a schematic structural diagram of a deep learning-based automatic processing system for unstructured data according to an embodiment of the present invention.

DETAILED DESCRIPTION

[0073] In order to make the purpose, technical solution and advantages of the embodiment of the present invention clearer, the technical solutions in the embodiments of the present invention will be clearly and completely described in combination with the accompanying drawings in the embodiments of the present invention. Definitely, the described embodiments are only a part of the embodiments of the present invention, not all of them. Based on the embodiments in the present invention, all other embodiments obtained by those of ordinary skill in the art without carrying out creative labor fall within the scope of protection of the present invention.

[0074] The technical solutions of the present invention are described in detail by specific embodiments below. The following specific embodiments can be combined with each other, and the same or similar concepts or processes may not be repeated in some embodiments.

[0075] FIG. 1 is a flowchart of a deep learning-based automatic processing method for unstructured data according to an embodiment of the present invention, as shown in FIG. 1, the method includes:

[0076] S101: acquiring unstructured data by an intelligent data collector; constructing a data feature fingerprint library; performing classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data; constructing a data quality assessment model by an incremental learning

algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; constructing a priority processing queue according to the scoring result, and performing data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set;

[0077] S102: performing deep learning processing on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and labelled data; constructing a feature association graph based on the target features; constructing a bidirectional inference channel on the feature association graph, the bidirectional inference channel including a forward channel and a reverse channel, the forward channel extracting numerical distribution features, time-order features and attribute features in the standardized training data set by a convolutional neural network so as to obtain dominant feature data, and the reverse channel extracting data item association rules, data change trends and data interaction modes in the standardized training data set through a recurrent neural network so as to obtain recessive feature data; performing feature combination on the dominant feature data and the recessive feature data by a dynamic weight allocation strategy to obtain a multi-dimensional feature tensor; and

[0078] S103: performing rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set; performing hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data; analyzing the initial structured data by a causal inference method to obtain a decision basis chain; generating optimization parameters based on the decision basis chain; and performing optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to obtain final structured data.

[0079] The intelligent data collector is a technical tool or device which can automatically acquire, process and store the unstructured data. It collects the unstructured data such as text, images, audio, and video by identifying and parsing the data of different sources, and ensures the integrity and accuracy of the data at the same time. The data feature fingerprint library is a database specially used for storing and managing the data features. Extraction and fingerprinting representation are performed on the multi-dimensional features (such as shape features, structure features, and statistical features) of data contents to form the unique identifiable data feature set for data

classification, retrieval and labelling. The feature association graph is a graph structure based data expression form used for describing an association relationship among the features. The nodes represent the features, and the edges represent dependency or interaction among the features, and thus an internal structure of the data can be better captured through the graph.

[0080] The bidirectional inference channel is a computation framework that combines forward inference and reverse inference; in the forward channel, it extracts the numerical distribution features, time-order features and attribute features by the convolutional neural network; and in the reverse channel, it captures the complex association rules and change trends between data items by the recurrent neural network. The dynamic weight allocation strategy is a technology for dynamically adjusting different feature weights according to task demands. By optimizing the feature combination process, the comprehensive analysis capability on the dominant and recessive features can be improved.

[0081] The multi-dimensional feature tensor is a tensor form for representing a high-dimensional data structure in deep learning. It integrates the dominant feature data and recessive feature data, thereby providing rich context information for subsequent inference and decision making.

[0082] In an optional embodiment, acquiring unstructured data by an intelligent data collector; constructing a data feature fingerprint library; performing classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data includes:

[0083] acquiring the unstructured data by the intelligent data collector, the intelligent data collector including a local feature channel and a global feature channel, the local feature channel performing word-level feature extraction and sentence-level feature extraction by a hierarchical attention network to obtain local association features, the global feature channel constructing a data relationship graph by a graph neural network and extracting global relationship features by message transfer, and performing attention fusion on the local association features and the global relationship features to obtain fused feature data;

[0084] performing multi-head attention computation on the fused feature data to obtain feature importance weights, screening the fused feature data according to the feature importance weights, mapping the screened features by a dynamic locality sensitive hashing algorithm to generate feature fingerprints, and constructing the feature fingerprints into the data feature fingerprint library that supports nearest neighbor query;

[0085] performing data enhancement on the fused feature data to obtain an enhanced feature set; computing contrast loss of sample pairs based on the enhanced feature set to obtain a semantic similarity matrix; constructing a data association graph through the nearest neighbor query result of the data feature fingerprint library and the semantic similarity matrix; transmitting known labels through graph attention on the data association graph to obtain an initial classification and labelling result, and extracting an entity relationship according to the data association graph to construct a domain knowledge graph;

[0086] constructing a meta-learner, performing assessment modeling on a classification result of each sample by taking a feature vector and confidence of the initial classification and labelling result as inputs to obtain a classification reliability score, tagging samples which are less than a preset reliability threshold and verified by the data feature fingerprint library as first-class to-be-re-checked data, performing logic verification on the initial classification and labelling result by the domain knowledge graph, and tagging samples which do not pass the verification as second-class to-be-re-checked data;

[0087] modeling a selection process of the first-class to-be-re-checked data and the second-class to-be-re-checked data as a Markov decision process, constructing a hybrid reward function according to the classification reliability score and a sample information gain, optimizing a sample selection strategy by deep reinforcement learning, and selecting an optimal to-be-re-checked sample for manual labelling to obtain a manual labelling result; and

[0088] performing weighted fusion on the manual labelling result and the initial classification and labelling result, performing consistency verification on the labelling result subjected to the weighted fusion by a similarity index of the data feature fingerprint library, and generating classified and labelled data containing the fusion feature data, the semantic similarity matrix, the classification reliability score and the domain knowledge graph verification result.

[0089] Exemplarily, when the intelligent data collector is utilized to acquire the unstructured data, the feature extraction is firstly carried out by the local feature channel and the global feature channel. In the local feature channel, the hierarchical attention network contains the word-level attention level and the sentence-level attention level. In the word-level attention level, it encodes an input text by a bidirectional long/short-term memory network to obtain a word vector sequence, computes the importance weight of each word through the attention mechanism, and performs weighting to obtain word-level feature representation. In the sentence-level attention level, it computes an association

degree among sentences based on the word-level features, generates sentence-level feature representation, and finally obtains the local association features.

[0090] In the global feature channel, the graph neural network firstly constructs a graph structure based on the association relationship among the data, the nodes represent data instances, and the edges represent relationships among instances. By a message transfer mechanism, each node acquires information from nearest neighbor nodes and updates own representation, and the node features containing global relationship information are obtained after multiple rounds of iteration. The local association features and the global relationship features are fused through the multi-head attention mechanism to obtain the fused feature data.

[0091] When the fused feature data is subjected to feature screening, the multi-head attention is computed to obtain the importance weights of the feature dimensions, and the feature dimensions with relatively high weights are reserved. The screened features are subjected to dimension reduction mapping by the dynamic locality sensitive hashing algorithm to generate compact feature fingerprints. The feature fingerprints are stored in an index structure that supports KNN query, and thus the data feature fingerprint library is constructed.

[0092] In the data enhancement stage, transformation such as random cutting and rotation is performed on the fused feature data to generate enhanced samples. The contrast loss of original samples and the enhanced samples is computed to obtain the semantic similarity matrix among the samples. The data association graph is constructed in combination with the nearest neighbor query result of the feature fingerprint library and the semantic similarity matrix. The known labels are transmitted on the association graph by a graph attention network to obtain the initial classification result of un-labelled samples. Meanwhile, an entity relationship triple is extracted from the association graph to construct the domain knowledge graph.

[0093] The meta-learner adopts a gradient-boosted tree model, the input features include the feature vectors and classification confidence of the samples, and the classification reliability score is outputted. The samples which have the score less than the threshold and are verified through the feature fingerprint library are tagged as the first-class to-be-re-checked data. The logic verification is performed on the initial classification result through a relationship rule in the domain knowledge graph, and the samples not meeting the rule are tagged as the second-class to-be-re-checked data.

[0094] The selection processes of the first-class to-be-re-checked data and second-class to-be-re-checked data are modelled based on the Markov decision process. In

the modeling process, each sample is treated as a state, and a decision of transferring the system from a current state to a next state is based on a series of constraints and objective functions. By modeling the selection process of the to-be-re-checked data, dynamic optimization can be realized, and thus the selected samples are ensured to be most valuable for manual labelling.

[0095] The classification reliability score reflects the credibility of the classification result of the samples, and it is obtained by comprehensive computation according to multi-dimensional factors such as consistency and confidence of the initial classification and labelling result and the sample features, and similarity with related samples in the data feature fingerprint library. The sample information gain refers to a potential contribution of the labelling result of each sample to the improvement of overall data classification model performance. By combining the classification reliability score and the sample information gain, the hybrid reward function is constructed, which aims to balance and select information samples with high values and avoid selection of repeated or low-efficiency samples.

[0096] The deep reinforcement learning is added to optimize the sample selection strategy. In this process, a reinforcement learning agent continuously updates the strategy through interaction with an environment, so as to realize maximization of the hybrid reward function. Specifically, the agent computes and selects an optimal to-be-rechecked sample according to current sample features, the classification reliability score and the information gain in each round of decision. Along with training, the agent will gradually learn to select the samples which are most valuable for improving the data quality and model performance more effectively.

[0097] After the manual labelling result is obtained, the manual labelling result and the initial classification and labelling result are subjected to weighted fusion. The weighted fusion realizes preferential trust on the manual labelling result by allocating the dynamic weights, and meanwhile, reliable parts in the initial classification and labelling result are reserved. This process ensures that the fused labelling result not only reflects high accuracy of manual labelling, but also maintains robustness of the model.

[0098] Through the similarity index of the data feature fingerprint library, the consistency verification is performed on the labelling result subjected to weighted fusion. The data feature fingerprint library provides high-dimensional feature fingerprints of the known samples, and by quick retrieval and similarity matching, whether the fused labelling result is consistent with the known samples can be verified. This step aims to troubleshoot potential labelling errors or inconsistency, which further improves the reliability of the

labelling result.

[0099] The finally generated classified labelling data not only contains the fused feature data, but also contains the semantic similarity matrix, the classification reliability score and the domain knowledge graph verification result. Such information is synthesized to form a complete set of data labelling system, which provides a powerful basis for subsequent data analysis, model training and decision support.

[0100] In this embodiment, through the dual-channel feature extraction mode of the hierarchical attention network and the graph neural network, the local semantic features of the data are captured, and moreover, the global relationship among the data is modeled, therefore, the comprehensiveness and accuracy of feature representation are improved. The feature fingerprint library is constructed by dynamic locality sensitive hashing and the nearest neighbor index, and in combination with comparative learning and graph attention transmission, efficient tag transmission is realized, thereby greatly reducing the manual labelling cost. Moreover, the logic verification is performed through the knowledge graph, which ensures the accuracy of the labelling result. Based on an active learning strategy of the deep reinforcement learning, the most valuable samples are adaptively selected for manual labelling, so that the labelling efficiency is remarkably improved. The consistency and reliability of the labelling result are ensured through a multi-verification mechanism, which improves the quality of classification and labelling.

[0101] In an optional embodiment, the constructing a data quality assessment model by an incremental learning algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; constructing a priority processing queue according to the scoring result, and performing data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set includes:

[0102] acquiring timestamp information of the classified and labelled data, performing computation by an exponential decay function to obtain a data time attenuation coefficient, and performing weighting processing on the classified and labelled data based on the data time attenuation coefficient to obtain a sample weight;

[0103] constructing the data quality assessment model, constructing a loss function by taking the sample weight as a weight item, optimizing the loss function by an online stochastic gradient descent algorithm until the model is converged, so as to obtain a trained data quality assessment model, and performing quality assessment on the classified and labelled data by the trained data quality assessment model;

[0104] computing an effective feature ratio based on the quality assessment result to obtain a data integrity score, obtaining a data timeliness score according to the data time attenuation coefficient, performing computation in combination with the sample weight to obtain a data value density score, and performing weighted fusion on the data integrity score, the data timeliness score and the data value density score to obtain a data quality score;

[0105] constructing the priority processing queue according to the data quality score, tagging the classified and labelled data with the data quality score less than a first preset threshold as to-be-removed data, tagging the classified and labelled data with the data quality score greater than a second preset threshold as priority processing data, computing the data processing priority based on the data quality score and the sample weight, and generating the priority data queue with processing tags;

[0106] according to the priority sequence of the priority data queue, carrying out unified transformation of field names and data formats on the classified and labelled data to obtain standardized features, constructing a time-order completion model based on the data timeliness score and the data integrity score, and completing missing fields by the time-order completion model to obtain structured features;

[0107] constructing a feature fingerprint according to the structured feature and the data value density score, identifying duplicated data based on the feature fingerprint, carrying out anomaly detection and de-noising processing in combination with the data quality score, and generating a quality processing record and cleaned feature data; and

[0108] carrying out standardized transformation on numerical features, class features and text features in the cleaned feature data respectively, guiding word vector mapping of the text features by the feature fingerprint, and combining the transformed standardized features with the data quality score, the data processing priority and the quality processing record to generate the standardized training data set.

[0109] Exemplarily, firstly, the classified and labelled data including data contents and timestamp information is acquired. For each piece of data, the time attenuation coefficient is computed based on its generation time. During specific implementation, a reference time point is set, for example, at current time, a time difference between each piece of data and the reference time is computed, and the attenuation coefficient is computed through the time difference. For example, for one piece of data generated 30 days ago, its time attenuation coefficient may be 0.85, and the data attenuation coefficient generated 60 days ago may be reduced to be 0.72.

[0110] Based on the computed time attenuation coefficient, weighting processing is

performed on the data to obtain the sample weight. The weight value reflects the timeliness importance degree of the data, and compared with new data, the data has a relatively high weight. For example, if the time attenuation coefficient of a certain piece of data is 0.85, its initial sample weight may be 0.85.

5 [0111] When constructing the data quality assessment model, the sample weight is treated as a weight item of the loss function. Model training is performed in the online stochastic gradient descent mode, a batch of data is stochastically selected for training each time, and the model parameters are continuously adjusted until convergence is achieved. The trained model can be used for assessing the data quality.

10 [0112] For each piece of data, scoring is performed from three dimensions of integrity, timeliness and value density, respectively:

 [0113] the integrity score is computed based on a proportion of effective features. For example, if a certain piece of data contains 10 feature fields, and 8 of the fields are effective, the integrity score may be 0.8.

15 [0114] The previously computed time attenuation coefficient is directly used as the timeliness score. For example, for the data with the attenuation coefficient of 0.85, the timeliness score is 0.85.

 [0115] The value density score is computed in combination with the sample weight. For example, the sample weight of a certain piece of data is 0.85, and its feature
20 information richness is relatively high, the value density score may be 0.88.

 [0116] The weighted fusion is performed on the scores of the three dimensions to obtain the final data quality score. For example, weights of 0.3, 0.3, and 0.4 can be given to the integrity score, the timeliness score and the value density score respectively, and a weighted average value is computed as the final score.

25 [0117] The priority processing queue is constructed based on the data quality score. Two thresholds are set, such as 0.6 and 0.8. The data with the score less than 0.6 is tagged as the data to be removed, and the data with the score greater than 0.8 is tagged as the data to be subjected to priority processing. The processing priority is computed in combination with the quality score and the sample weight to generate the priority queue.

30 [0118] The data is processed according to the priority sequence. Firstly, field unification is performed to unify dates in different formats into a standard format and transform numerical values of different units into a unified unit. For missing fields, filling is performed based on the time-order completion model. For example, for missing sales volume data, prediction filling can be performed based on historical data.

35 [0119] The feature fingerprints are constructed for duplication eliminating and

exception detection. The feature fingerprints can contain a hash value combination of the key fields. For example, for commodity data, fields of categories, specifications, production dates and the like can be combined to generate the fingerprints. Based on the fingerprints, the duplicated data can be identified; and the exception detection is performed
 5 in combination with the quality score.

[0120] Finally, the standardized processing is performed on the cleaned data. Normalization is performed on the numerical features, the class features are transformed into an encoding form, and the text features are transformed into word vectors. The standardized features are combined with the information such as the quality score to
 10 generate the final training data set.

[0121] In this embodiment, by the time attenuation mechanism and the multi-dimensional quality assessment, accurate quantification of the data values is realized, thus the accuracy and efficiency of data processing are improved, and the data quality assessment is more objective and reasonable. The priority queue and the feature fingerprint
 15 mechanism are adopted, so that intelligent scheduling and exception identification in the data processing can be realized, the resource consumption in the data processing is reduced, and the efficiency and accuracy of data cleaning are improved. By the standardized processing and feature transformation, the uniformity of the data formats and the standardability of the feature expression are ensured, the effect of subsequent model
 20 training is improved, and the reliability and stability of data application are enhanced.

[0122] In an optional embodiment, the performing deep learning processing on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and
 25 labelled data includes:

[0123] constructing the multi-level cascade network structure, the multi-level cascade network structure including a feature extraction level, an attention level and a feature fusion level, the feature extraction level consisting of a plurality of convolutional blocks, the attention level adopting a self-attention mechanism, and the feature fusion level being
 30 used for feature dimension reduction and combination;

[0124] performing class information extraction on the classified and labelled data in the standardized training data set, constructing a class embedding matrix, and performing computation based on the class embedding matrix and the standardized training data set to obtain a class mapping matrix;

[0125] processing the standardized training data set by the feature extraction level of
 35

the multi-level cascade network structure, performing product operation on the class mapping matrix and extracted features to obtain class perception features, and splicing the class perception features extracted at different levels in a channel dimension to obtain multi-scale cascade features;

5 [0126] in the attention level, generating a query matrix, a key matrix and a value matrix based on the multi-scale cascade features and the class embedding matrix, computing the similarity of the query matrix and the key matrix through the self-attention mechanism to obtain an attention weight, and multiplying the attention weight by the value matrix to obtain attention features;

10 [0127] in the feature fusion level, performing residual connection and normalization processing on the attention features to obtain standardized features, constructing a gating weight based on the standardized features and the class embedding matrix, and performing weighting on the gating weight and the standardized features to obtain enhanced features; and

15 [0128] transmitting the enhanced features of each level by layer in the multi-level cascade network structure and performing feature cascade, and performing cross-layer feature aggregation to obtain the target features in the standardized training data set.

 [0129] Exemplarily, in the construction process of the multi-level cascade network structure, it is firstly needed to design a feature extraction level. The feature extraction level
20 consists of a plurality of convolutional blocks, each convolutional block contains three convolutional layers, the sizes of convolution kernels are 3×3 , 5×5 and 7×7 respectively, and the step length is 1; and the filling mode is a same filling mode. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function, and is used for extracting feature information at different scales.

25 [0130] For the classified and labelled data in the standardized training data set, the class information is transformed into a vector form by a one-hot encoding mode. If the data set contains 10 classes, each class tag is transformed into a 10-dimensional vector. These vectors are combined into the class embedding matrix, and the matrix dimension is class number $\times$ embedding dimension. By taking an image classification task as an example, the
30 input image size is $224 \times 224 \times 3$, and the size of the feature graph obtained from the feature extraction level will be $28 \times 28 \times 256$.

 [0131] In the feature extraction process, the class mapping matrix and the extracted features are subjected to dot product operation to obtain features with class perception capability. For the features extracted by different convolutional blocks, splicing is
35 performed in a channel dimension to form the multi-scale cascade features. By taking the

above image classification task as an example, three feature dimensions of different scales are $28 \times 28 \times 256$, $14 \times 14 \times 512$ and $7 \times 7 \times 1024$ respectively, and the multi-scale features of $28 \times 28 \times 1792$ are obtained after splicing.

[0132] In the processing process of the attention level, the query matrix is generated based on the multi-scale cascade features, and the key matrix and the value matrix are generated based on the class embedding matrix. The similarity between the query matrix and the key matrix is computed to obtain the attention weights, and the weights are normalized by a softmax function. The normalized attention weights are multiplied by the value matrix to obtain the attention features fused with the class information.

[0133] In the feature fusion level, residual connection is firstly performed on the attention features, the input features and the attention features are added, and then standardized processing is performed by level normalization. Based on the standardized features and the class embedding matrix, a fully connected layer is used for constructing the gating weight mechanism, and this weight is used for controlling the importance of different features. The gating weights and the standardized features are multiplied element by element to obtain the enhanced feature representation.

[0134] Finally, the enhanced features of each level are transmitted level by level in the network, and feature aggregation is performed in the cross-level connection mode. Specifically, low-level features are aligned with high-level features in size by up-sampling, and then splicing is performed in the channel dimension to finally obtain target features fusing the multi-level information.

[0135] In this embodiment, the multi-scale feature information in the data can be effectively captured through the multi-level cascade structure in combination with the self-attention mechanism, thus the comprehensiveness and accuracy of feature extraction can be improved and the adaptability of the model to targets of different scales can be enhanced. The class information is added to guide the feature extraction process, and through the construction of the class perception features and the application of the attention mechanism, the identification capability of the model to the class-related features can be improved and the noise interference in the feature extraction process can be reduced. A mode of combining the gating mechanism and residual connection is adopted, which realizes adaptive fusion and optimization of the features, improves the discrimination of feature representation, maintains the original feature information, and enhances the generalization capability of the model.

[0136] In an optional embodiment, the performing rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an

inference rule set includes:

[0137] performing dimension reduction processing on the multi-dimensional feature tensor by a feature dimension reduction network to obtain a feature sequence, constructing a feature index tree on the feature sequence, setting a dynamically sized sliding window based on the feature index tree, adaptively adjusting the size of the sliding window according to the feature sequence, and performing sliding scanning on the feature sequence to extract a feature mode;

[0138] inputting the feature mode into a pre-training text encoder to obtain a feature encoding vector, and inputting a rule mode in the preset domain knowledge base into the pre-training text encoder to obtain a rule encoding vector; computing cosine similarity between the feature encoding vector and the rule encoding vector as a matching degree;

[0139] computing a mean value and a standard deviation of the matching degree, determining an adaptive threshold coefficient according to a preset confidence interval, performing weighting on the mean value and the standard deviation based on the adaptive threshold coefficient to obtain an adaptive threshold, screening the rule modes with the matching degrees greater than the adaptive threshold, and forming a candidate rule set by the screened rule modes and corresponding structured templates; and

[0140] performing dependency analysis on rules in the candidate rule set, establishing a dependency relationship among the rules based on input features and output features of the rules, representing the dependency relationship as a directed graph, nodes representing the rules and directed edges representing the dependency relationship, computing an in-degree and an out-degree of each rule node, determining a dependency level of the rules, and allocating the rules having the direct dependency relationship to adjacent levels by an improved topological sorting algorithm, so as to obtain the inference rule set.

[0141] Exemplarily, firstly, the feature dimension reduction processing is performed on the inputted multi-dimensional feature tensor. The deep neural network is treated as the feature dimension reduction network, which contains a multi-level convolutional layer and a pooling layer. By taking a medical diagnosis scenario as an example, the inputted multi-dimensional feature tensor contains various examination indexes of a patient, and the dimension is 128 x 64 x 32. The inputted multi-dimensional feature tensor is transformed into a one-dimensional feature sequence by the feature dimension reduction network, and the length is 512.

[0142] After the feature sequence is obtained, a feature index tree is constructed. The feature sequence is stored by a B + tree structure, the feature values are stored in leaf nodes and the index information is stored in non-leaf nodes. The dynamically size sliding window

is set on the feature sequence, and the initial window size is set as 32, and dynamic adjustment is carried out according to a local change condition of the feature sequence. When the feature sequence fluctuates relatively greatly, the size of the window is reduced so as to capture fine-grained features; and when the feature sequence is stable, the size of the window can be increased so as to obtain macro features.

[0143] The sliding window scans the feature sequence so as to extract the feature modes. Each feature mode contains a feature value sequence in the window. The extracted feature modes are inputted into a pre-trained BERT text encoder so as to obtain a 768-dimensional feature encoding vector. Meanwhile, the rule modes in the preset domain knowledge base are also inputted into the same encoder so as to obtain the rule encoding vectors.

[0144] The cosine similarity between the feature encoding vector and the rule encoding vectors is computed so as to obtain the matching degree. By taking a medical diagnosis rule as an example, there is a certain rule mode that “Blood pressure continues to rise and is accompanied by headaches”, and the cosine similarity between the rule encoding vector and the encoding vector of a certain feature mode is 0.85.

[0145] Statistics is performed on the mean values and standard deviations of all the matching degrees. If the mean values are 0.7, the standard deviations will be 0.15. A 95% confidence interval is set, and a corresponding adaptive threshold coefficient is 1.96. The mean values and the standard deviations are weighted to obtain an adaptive threshold: $0.7 + 1.96 \times 0.15 = 0.994$. The rule modes with the matching degrees greater than the threshold are screened to form the candidate rule sets.

[0146] The dependency relationship analysis is performed on the rules in the candidate rule sets. By analyzing the input features and output features of the rules, the dependency relationship network among the rules can be constructed. The dependency relationship is represented by a directed graph, the nodes represent the rules, and the directed edges represent dependency among the rules. The in-degree and an out-degree of each node are computed, thereby identifying relative positions and importance of the rules.

[0147] The rules are hierarchically divided by the improved topological sorting algorithm according to the dependency relationship among the rules. The topological sorting result is an ordered list, so that an executive sequence of the rules is ensured to conform to logic dependency, and meanwhile, cyclic dependency is avoided.

[0148] Finally, the sorted rule set is arranged into an inference rule set. The rules in the inference rule set are executed step by step according to the dependency hierarchies, so that the orderliness and high efficiency of the inference process are ensured.

[0149] In this embodiment, the dynamically adjusted sliding window is used for extracting the feature mores, so that the feature information of different scales can be adaptively captured, and the flexibility and accuracy of feature extraction can be improved. The similarity between the feature mode and the rule mode is computed by a pre-training
 5 text encoder, which fully utilizes the semantic understanding capability of the deep learning model, thereby improving the accuracy and generalization capability of rule matching. The hierarchical inference rule set is constructed based on the rule dependency relationship, which makes the logic relationship among the rules clearer, thereby improving the interpretability and inference efficiency of the inference process.

10 [0150] In an optional embodiment, the performing hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data includes:

[0151] extracting the feature vectors of rule-related regions from the feature sequence for the rules at each rule level in the inference rule set, inputting the feature vectors into a
 15 rule matching network, the rule matching network including a feature transformation layer and a matching degree computation layer, outputting rule applicability scores, and standardizing the rule applicability scores to serve as rule confidence;

[0152] performing descending on the rules at each rule level based on the rule confidence to obtain a sorting rule set, applying the rules to the corresponding structured
 20 templates in sequence according to the sorting sequence, and generating structured fragments by feature mapping, the structured templates defining a mapping relationship between the features and the structured data;

[0153] detecting action regions of the rules in the sorting rule set, when the action regions of a plurality of rules are overlapped, extracting a range of an overlapped region,
 25 computing a coverage degree of each rule in the overlapped region, taking a weighted sum of the rule confidence and the coverage degrees as a rule priority, and selecting the rule with the highest priority to be applied to the overlapped region, thereby ensuring to generate a unique structured fragment;

[0154] constructing a multi-level verification network containing an attribute
 30 verification level, a relationship verification level and a value domain verification level, verifying the integrity of the structured fragment based on an attribute graph in the attribute verification level, verifying the relationship reasonability between the structured fragments based on a knowledge graph in the relationship verification level, verifying the validity of attribute values in the structured fragments based on a statistical model in the value domain
 35 verification level, and integrating verification result of each level to obtain verified

structured fragments; and

[0155] extracting semantic features in the verified structured fragments, computing the semantic similarity among the fragments, constructing hierarchical organization structures of the fragments based on the semantic similarity, and integrating the structured fragments related to semantics according to the hierarchical organization structures to generate the initial structured data meeting a preset specification.

[0156] Exemplarily, the hierarchical processing is performed on the multi-dimensional feature tensor through the inference rule set so as to generate the initial structured data. Firstly, regional extraction is performed on the feature sequence according to the rule levels in the inference rule set. In the extraction process, for the rules in each rule level, the feature regions related to rule conditions are identified, and the feature vectors thereof are extracted to serve as inputs.

[0157] The extracted feature vectors are inputted into the rule matching network. The rule matching network consists of a feature transformation layer and a matching degree computation layer; in the feature transformation layer, normalization and feature enhancement are performed on the inputted feature vectors, and in the matching degree computation layer, an applicability score of each rule is outputted by computing the matching degree of the features and the rule conditions. The applicability score reflects the matching degree between the rules and the features. Then, standardized processing is performed on the applicability scores to generate the rule confidence, and the confidence serves as key indexes for measuring the reliability of the rules.

[0158] The rules in each rule level are sorted by descending based on the rule confidence so as to generate the sorting rule set. In the sorting rule set, it is sequentially arranged according to the importance and applicability of the rules, thereby ensuring that the high-confidence-degree rules are preferentially applied. Then, the rules in the sorting rule set are sequentially applied to corresponding structured templates. The structured templates define the mapping rules from the features to the structured data, and the structured fragments are generated in combination with the mapping relationships between the feature vectors and the templates. The structured fragments are basic units of the initial structured data.

[0159] When the action regions of the rules are overlapped, the range of the overlapped region is detected, all the rules in the overlapped region are extracted, and the coverage degree of each rule in the overlapped region is computed. The coverage degree represents a range proportion of the action of the rules in the overlapped region. The rule confidence and the coverage degree are weighted to obtain the rule priority. The rule with

the highest priority is applied to the overlapped region, so that it is ensured that only the unique structured fragment is generated in each region, and data conflicts and redundancy are avoided.

[0160] The multi-level verification network is constructed to perform
 5 multi-dimensional verification on the generated structured fragments. The multi-level verification network includes an attribute verification level, a relationship verification level and a value domain verification level. In the attribute verification level, it checks the integrity of the structured fragments based on the attribute graph, for example, it confirms whether the fragment contains all necessary attribute values. In the relationship verification
 10 level, it verifies whether the relationships among the structured fragments are reasonable based on the knowledge graph, for example, it checks whether different fragments conform to predefined logic or semantic association. In the value domain verification level, it checks the validity of the fragment attribute values based on the statistical model, for example, it confirms whether numerical values are in a reasonable range. All the verification results are
 15 integrated, and the fragments not conforming to specifications are corrected or removed, thus obtaining the final verified structured fragments.

[0161] The semantic features are extracted from the verified structured fragments, and the relevance among the fragments is computed through the semantic similarity. The semantic features capture deep semantic information of fragment contents, and the
 20 semantic similarity reflects the semantic association degree among the fragments. The hierarchical organization structure of the fragments is constructed based on the semantic similarity, and the semantically associated fragments are aggregated into larger units. By integrating the associated fragments, the initial structured data conforming to preset specifications is generated.

[0162] In this embodiment, by hierarchical processing and rule matching, the accuracy
 25 of feature extraction can be improved; the rule matching network can be adaptive to feature representation of different scenarios, thus improving the flexibility and robustness of rule application. The multi-dimensional data quality control is realized by the multi-level verification network, and therefore, the accuracy of the structured result is ensured from the
 30 attribute integrity, relationship rationality, value effectiveness and the like, and the data reliability is remarkably improved. The intelligent integration of the structured fragments is realized based on the hierarchical organization of the semantic similarity, so that the problems of data redundancy and inconsistency are solved and the normalization and availability of structured data are improved.

[0163] In an optional embodiment, the analyzing the initial structured data by a
 35

causal inference method to obtain a decision basis chain; generating optimization parameters based on the decision basis chain; and performing optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to obtain final structured data includes:

5 [0164] computing a partial association coefficient for a variable pair in the initial structured data, performing a conditional independence test based on the partial association coefficient to obtain a conditional independence relationship among variables, and constructing a causality graph according to the conditional independence relationship;

10 [0165] constructing a linear structure equation model based on the causality graph, optimizing parameters of the linear structure equation model by maximum likelihood estimation to obtain a structural causal model, the structural causal model being used for quantifying causal effect intensity among the variables in the causality graph;

15 [0166] computing the number of associated edges of each node in the causality graph, determining the nodes with the number of the associated edges exceeding a preset associated edge number threshold as decision nodes, tracing associated nodes level by level from the decision nodes along causality edges in the causality graph to obtain factor chains, computing a cumulative causal effect of the factor chains based on the structural causal model, and determining the factor chains with the cumulative causal effect exceeding a preset effect threshold as candidate decision basis chains;

20 [0167] in the structural causal model, changing the value of each node in the candidate decision basis chains and substituting into the structural causal model to obtain a corresponding downstream node predicted value after the value of the node is changed, determining the stability of the candidate decision basis chains based on a change amplitude of the predicted value, and determining the candidate decision basis chains with the stability meeting a preset condition as decision basis chains;

25 [0168] computing an influence degree of each node in the decision basis chains on a target variable based on the structural causal model, screening nodes with the influence degree exceeding a preset influence threshold as a feature set according to the influence degree, taking the influence degree of each node in the feature set as a feature importance score, and combining the feature importance score with the class discrimination of the classified and labelled data to obtain a feature optimization parameter;

30 [0169] performing weighted adjustment on the value of the corresponding feature in the initial structured data based on the feature optimization parameter to obtain optimized data, the feature optimization parameter being used for determining a weight coefficient of feature adjustment; and

35

[0170] computing the feature distribution difference between the optimized data and the initial structured data to obtain an optimization effect index; inputting the optimized data into the structural causal model to verify causal relationship constraint so as to obtain a verification result, performing iterative optimization on the optimized data according to the optimization effect index and the verification result, and when the optimization index reaches a preset target and the verification result meets causal consistency, determining the current optimized data as the final structured data.

[0171] Exemplarily, all variables in the initial structured data are paired and the partial association coefficients between each pair of variables are computed. The partial association coefficients reflect the degree of a linear relationship between two variables and eliminate the interference influence of other variables. By analyzing the partial association coefficients between the variables, whether there is a direct or indirect causal relationship possibly between the variables can be preliminarily determined.

[0172] According to the computed partial association coefficients, the conditional independence test can be further executed for determining whether the variables are independent after other variables are controlled. If the conditions of the two variables are independent, there is no causal relationships between the two variables. Based on the conditional independence relationships, the causal graph representing the causal relationships between the variables is constructed. The nodes in the causal graph represent the variables and the directed edges represent direct causal effects between the variables. In the construction process, it is needed to be ensured that the structure of the graph is an acyclic graph and meets priori knowledge or domain constraints.

[0173] The structural equation model is constructed based on the causal graph and used for quantifying the strength of the causal relationships among the variables. In the structural equation model, the value of each variable can be represented as a linear combination of causal precursor variables of the each variable. The model parameters are optimized by a maximum likelihood estimation method, and finally the structural causal model describing a quantitative relationship of the causal graph is generated. The model is used for evaluating the range of the causal effect among the variables, thus providing a quantitative basis for subsequent analysis.

[0174] The nodes in the causal graph are analyzed, and the nodes with the associated edge quantity exceeding the preset threshold are tagged as the decision nodes. The decision nodes are generally at key positions in the causal graph and have relatively high influence on the overall causal structure. Other associated nodes are traced back step by step along a causal path from each decision node, so as to form a series of causal chains. The

accumulated causal effects of the causal chains are quantified by the structural causal model, and the accumulated causal effects represent overall causal influence strength from a starting point to an end point of the chains. By setting a threshold of the accumulated causal effects, the chains with relatively strong causal effects are screened as the candidate decision basis chains.

[0175] The stability of the candidate decision basis chains is verified. The values of the nodes in the chains are changed one by one in the structural causal model, and the change amplitude of the predicted values of downstream nodes in the chains is observed. If a certain candidate decision basis chain can maintain relatively high prediction stability when the value of the node changes, its reliability is higher. The candidate decision basis chain meeting a stability condition is confirmed as the final decision basis chain.

[0176] The direct or indirect influence of each node in the decision basis chain on a target variable is computed by the structural causal model, and is quantified into the influence degree. The nodes with the influence degree exceeding the threshold are screened as the feature set, and meanwhile, the influence degree of each node is recorded as the feature importance score. In combination with statistical characteristics of the class discrimination in the classified and labelled data, the feature importance score and the class discrimination are comprehensively computed to generate the feature optimization parameters used in the optimization process.

[0177] Based on the feature optimization parameters, weighted adjustment is performed on the corresponding features in the initial structured data. In the weighted adjustment process, it takes into account the sizes of the feature optimization parameters, it is ensured that higher weights are given to the key features, thereby performing more remarkable optimization on the values of the key features. The adjusted data is used as the optimized data and needs further verification on its effectiveness.

[0178] The optimized data is compared with the initial data, and the feature distribution change condition of the optimized data is analyzed to generate the optimization effect index. Meanwhile, the optimized data is re-inputted into the structural causal model for verifying whether the optimized data meets the causal relationship constraint. If the causal relationship constraint is destroyed in the optimization process, it is needed to re-adjust the feature optimization parameters and the data weight allocation strategy.

[0179] Through multiple rounds of iteration, the consistency between feature distribution and target distribution of the optimized data is gradually improved, and moreover, the optimization result maintains causal consistency. When the optimization effect index reaches a preset target range and the causal verification result meets the

constraint requirement, the iteration will be stopped, and the current optimized data is determined as the final structured data.

[0180] The finally generated structured data meets preset requirements in the causal consistency, feature distribution rationality and optimization effect, and can be directly
5 used for subsequent data analysis or decision support tasks.

[0181] In this embodiment, the causal graph and the structural causal model are constructed, and the causal relationship in the data is deeply explored, thereby improving the accuracy and interpretability of data analysis. Compared with conventional association analysis, the causal inference can better reflect the real relationship among the variables,
10 and avoid misleading caused by false association. The concept of the decision basis chain is added, and the reliable theoretical basis is provided for data optimization by tracking the causal chain and assessing the stability of the causal chain. According to the causality-based feature selection and optimization method, the key features can be identified more effectively and the generalization capability and robustness of the model
15 are improved. The causal consistency is taken into account in the optimization process, and the optimized data is ensured to still meet the original causal relationship constraints by iterative verification. According to the method, the data quality is improved, moreover, the interpretability and credibility of the optimization result are ensured, and thus a more reliable data basis is provided for subsequent decision analysis and model application.

[0182] FIG. 2 is a schematic structural diagram of a deep learning-based automatic processing system for unstructured data according to an embodiment of the present invention, as shown in FIG. 2, the system includes:

[0183] a first unit, configured to acquire unstructured data by an intelligent data collector; construct a data feature fingerprint library; perform classification and labelling on
25 the unstructured data based on the data feature fingerprint library to obtain classified and labelled data; construct a data quality assessment model by an incremental learning algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; construct a priority processing queue according to the scoring result, and perform data
30 cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set;

[0184] a second unit, configured to perform deep learning processing on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set
35 through a self-attention mechanism according to class information of the classified and

labelled data; construct a feature association graph based on the target features; construct a bidirectional inference channel on the feature association graph, the bidirectional inference channel including a forward channel and a reverse channel, the forward channel extracting numerical distribution features, time-order features and attribute features in the standardized training data set by a convolutional neural network so as to obtain dominant feature data, and the reverse channel extracting data item association rules, data change trends and data interaction modes in the standardized training data set through a recurrent neural network so as to obtain recessive feature data; and perform feature combination on the dominant feature data and the recessive feature data by a dynamic weight allocation strategy to obtain a multi-dimensional feature tensor; and

[0185] a third unit, configured to perform rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set; perform hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data; analyze the initial structured data by a causal inference method to obtain a decision basis chain; generate optimization parameters based on the decision basis chain; and perform optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to obtain final structured data.

[0186] In a third aspect,

[0187] an electronic device provided by the embodiment of the present invention includes:

[0188] a processor; and

[0189] a memory configured to store processor executable instructions;

[0190] the processor is configured to call instructions stored in the memory to execute the abovementioned method.

[0191] In a fourth aspect,

[0192] the embodiment of the present invention provides a computer-readable storage medium having computer program instructions stored therein, when executed by a processor, the computer program instructions execute the abovementioned method.

[0193] The present invention can be a method, an apparatus, a system, and/or a computer program product. The computer program product can include a computer-readable storage medium that contains computer-readable program instructions for executing the present invention in various aspects.

[0194] Finally, it is to be noted that: the above embodiments are only used for illustrating the technical solution of the present invention, not to limit it. Although the

present invention is described in detail with reference to the aforementioned embodiments, those of ordinary skill in the art should understand that it may still modify the technical solutions recorded in the aforementioned embodiments, or replace some or all of the technical features thereof. These modifications or substitutions do not make the essence of the corresponding technical solution out of the scope of the technical solutions of each embodiment of the present invention.

CLAIMS

We claim:

1. A deep learning-based automatic processing method for unstructured data, comprising:

5 acquiring unstructured data by an intelligent data collector; constructing a data feature fingerprint library; performing classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data; constructing a data quality assessment model by an incremental learning algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; constructing a priority
10 processing queue according to the scoring result, and performing data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set;

 performing deep learning processing on the standardized training data set by a
15 multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and labelled data; constructing a feature association graph based on the target features; constructing a bidirectional inference channel on the feature association graph, the bidirectional inference channel including a
20 forward channel and a reverse channel, the forward channel extracting numerical distribution features, time-order features and attribute features in the standardized training data set by a convolutional neural network so as to obtain dominant feature data, and the reverse channel extracting data item association rules, data change trends and data interaction modes in the standardized training data set through a recurrent neural network
25 so as to obtain recessive feature data; performing feature combination on the dominant feature data and the recessive feature data by a dynamic weight allocation strategy to obtain a multi-dimensional feature tensor; and

 performing rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set; performing hierarchical processing
30 on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data; analyzing the initial structured data by a causal inference method to obtain a decision basis chain; generating optimization parameters based on the decision basis chain; and performing optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to
35 obtain final structured data.

2. The method according to claim 1, wherein the acquiring unstructured data by an intelligent data collector; constructing a data feature fingerprint library; performing classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data comprises:

5 acquiring the unstructured data by the intelligent data collector, the intelligent data collector comprising a local feature channel and a global feature channel, the local feature channel performing word-level feature extraction and sentence-level feature extraction by a hierarchical attention network to obtain local association features, the global feature channel constructing a data relationship graph by a graph neural network and extracting
10 global relationship features by message transfer, and performing attention fusion on the local association features and the global relationship features to obtain fused feature data;

 performing multi-head attention computation on the fused feature data to obtain feature importance weights, screening the fused feature data according to the feature importance weights, mapping the screened features by a dynamic locality sensitive hashing
15 algorithm to generate feature fingerprints, and constructing the feature fingerprints into the data feature fingerprint library that supports nearest neighbor query;

 performing data enhancement on the fused feature data to obtain an enhanced feature set; computing contrast loss of sample pairs based on the enhanced feature set to obtain a semantic similarity matrix; constructing a data association graph through the nearest
20 neighbor query result of the data feature fingerprint library and the semantic similarity matrix; transmitting known labels through graph attention on the data association graph to obtain an initial classification and labelling result, and extracting an entity relationship according to the data association graph to construct a domain knowledge graph;

 constructing a meta-learner, performing assessment modeling on a classification result
25 of each sample by taking a feature vector and confidence of the initial classification and labelling result as inputs to obtain a classification reliability score, tagging samples which are less than a preset reliability threshold and verified by the data feature fingerprint library as first-class to-be-re-checked data, performing logic verification on the initial classification and labelling result by the domain knowledge graph, and tagging samples
30 which do not pass the verification as second-class to-be-re-checked data;

 modeling a selection process of the first-class to-be-re-checked data and the second-class to-be-re-checked data as a Markov decision process, constructing a hybrid reward function according to the classification reliability score and a sample information gain, optimizing a sample selection strategy by deep reinforcement learning, and selecting
35 an optimal to-be-re-checked sample for manual labelling to obtain a manual labelling result;

and

performing weighted fusion on the manual labelling result and the initial classification and labelling result, performing consistency verification on the labelling result subjected to the weighted fusion by a similarity index of the data feature fingerprint library, and
 5 generating classified and labelled data containing the fusion feature data, the semantic similarity matrix, the classification reliability score and the domain knowledge graph verification result.

3. The method according to claim 1, wherein the constructing a data quality assessment model by an incremental learning algorithm, the data quality assessment model
 10 scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; constructing a priority processing queue according to the scoring result, and performing data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set comprises:

15 acquiring timestamp information of the classified and labelled data, performing computation by an exponential decay function to obtain a data time attenuation coefficient, and performing weighting processing on the classified and labelled data based on the data time attenuation coefficient to obtain a sample weight;

constructing the data quality assessment model, constructing a loss function by taking
 20 the sample weight as a weight item, optimizing the loss function by an online stochastic gradient descent algorithm until the model is converged, so as to obtain a trained data quality assessment model, and performing quality assessment on the classified and labelled data by the trained data quality assessment model;

computing an effective feature ratio based on the quality assessment result to obtain a
 25 data integrity score, obtaining a data timeliness score according to the data time attenuation coefficient, performing computation in combination with the sample weight to obtain a data value density score, and performing weighted fusion on the data integrity score, the data timeliness score and the data value density score to obtain a data quality score;

constructing the priority processing queue according to the data quality score, tagging
 30 the classified and labelled data with the data quality score less than a first preset threshold as to-be-removed data, tagging the classified and labelled data with the data quality score greater than a second preset threshold as priority processing data, computing the data processing priority based on the data quality score and the sample weight, and generating the priority data queue with processing tags;

35 according to the priority sequence of the priority data queue, carrying out unified

transformation of field names and data formats on the classified and labelled data to obtain standardized features, constructing a time-order completion model based on the data timeliness score and the data integrity score, and completing missing fields by the time-order completion model to obtain structured features;

5 constructing a feature fingerprint according to the structured feature and the data value density score, identifying duplicated data based on the feature fingerprint, carrying out anomaly detection and de-noising processing in combination with the data quality score, and generating a quality processing record and cleaned feature data; and

 carrying out standardized transformation on numerical features, class features and text
10 features in the cleaned feature data respectively, guiding word vector mapping of the text features by the feature fingerprint, and combining the transformed standardized features with the data quality score, the data processing priority and the quality processing record to generate the standardized training data set.

 4. The method according to claim 1, wherein the performing deep learning processing
15 on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and labelled data comprises:

 constructing the multi-level cascade network structure, the multi-level cascade
20 network structure comprising a feature extraction level, an attention level and a feature fusion level, the feature extraction level consisting of a plurality of convolutional blocks, the attention level adopting a self-attention mechanism, and the feature fusion level being used for feature dimension reduction and combination;

 performing class information extraction on the classified and labelled data in the
25 standardized training data set, constructing a class embedding matrix, and performing computation based on the class embedding matrix and the standardized training data set to obtain a class mapping matrix;

 processing the standardized training data set by the feature extraction level of the
30 multi-level cascade network structure, performing product operation on the class mapping matrix and extracted features to obtain class perception features, and splicing the class perception features extracted at different levels in a channel dimension to obtain multi-scale cascade features;

 in the attention level, generating a query matrix, a key matrix and a value matrix based
35 on the multi-scale cascade features and the class embedding matrix, computing the similarity of the query matrix and the key matrix through the self-attention mechanism to

obtain an attention weight, and multiplying the attention weight by the value matrix to obtain attention features;

in the feature fusion level, performing residual connection and normalization processing on the attention features to obtain standardized features, constructing a gating weight based on the standardized features and the class embedding matrix, and performing weighting on the gating weight and the standardized features to obtain enhanced features; and

transmitting the enhanced features of each level by layer in the multi-level cascade network structure and performing feature cascade, and performing cross-layer feature aggregation to obtain the target features in the standardized training data set.

5. The method according to claim 1, wherein the performing rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set comprises:

performing dimension reduction processing on the multi-dimensional feature tensor by a feature dimension reduction network to obtain a feature sequence, constructing a feature index tree on the feature sequence, setting a dynamically sized sliding window based on the feature index tree, adaptively adjusting the size of the sliding window according to the feature sequence, and performing sliding scanning on the feature sequence to extract a feature mode;

inputting the feature mode into a pre-training text encoder to obtain a feature encoding vector, and inputting a rule mode in the preset domain knowledge base into the pre-training text encoder to obtain a rule encoding vector; computing cosine similarity between the feature encoding vector and the rule encoding vector as a matching degree;

computing a mean value and a standard deviation of the matching degree, determining an adaptive threshold coefficient according to a preset confidence interval, performing weighting on the mean value and the standard deviation based on the adaptive threshold coefficient to obtain an adaptive threshold, screening the rule modes with the matching degrees greater than the adaptive threshold, and forming a candidate rule set by the screened rule modes and corresponding structured templates; and

performing dependency analysis on rules in the candidate rule set, establishing a dependency relationship among the rules based on input features and output features of the rules, representing the dependency relationship as a directed graph, nodes representing the rules and directed edges representing the dependency relationship, computing an in-degree and an out-degree of each rule node, determining a dependency level of the rules, and allocating the rules having the direct dependency relationship to adjacent levels by an

improved topological sorting algorithm, so as to obtain the inference rule set.

6. The method according to claim 1, wherein the performing hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data comprises:

5 extracting the feature vectors of rule-related regions from the feature sequence for the rules at each rule level in the inference rule set, inputting the feature vectors into a rule matching network, the rule matching network comprising a feature transformation layer and a matching degree computation layer, outputting rule applicability scores, and standardizing the rule applicability scores to serve as rule confidence;

10 performing descending on the rules at each rule level based on the rule confidence to obtain a sorting rule set, applying the rules to the corresponding structured templates in sequence according to the sorting sequence, and generating structured fragments by feature mapping, the structured templates defining a mapping relationship between the features and the structured data;

15 detecting action regions of the rules in the sorting rule set, when the action regions of a plurality of rules are overlapped, extracting a range of an overlapped region, computing a coverage degree of each rule in the overlapped region, taking a weighted sum of the rule confidence and the coverage degrees as a rule priority, and selecting the rule with the highest priority to be applied to the overlapped region, thereby ensuring to generate a
20 unique structured fragment;

 constructing a multi-level verification network containing an attribute verification level, a relationship verification level and a value domain verification level, verifying the integrity of the structured fragment based on an attribute graph in the attribute verification level, verifying the relationship reasonability between the structured fragments based on a
25 knowledge graph in the relationship verification level, verifying the validity of attribute values in the structured fragments based on a statistical model in the value domain verification level, and integrating verification result of each level to obtain verified structured fragments; and

 extracting semantic features in the verified structured fragments, computing the
30 semantic similarity among the fragments, constructing hierarchical organization structures of the fragments based on the semantic similarity, and integrating the structured fragments related to semantics according to the hierarchical organization structures to generate the initial structured data meeting a preset specification.

7. The method according to claim 1, wherein the analyzing the initial structured data
35 by a causal inference method to obtain a decision basis chain; generating optimization

parameters based on the decision basis chain; and performing optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to obtain final structured data comprises:

5 computing a partial association coefficient for a variable pair in the initial structured data, performing a conditional independence test based on the partial association coefficient to obtain a conditional independence relationship among variables, and constructing a causality graph according to the conditional independence relationship;

10 constructing a linear structure equation model based on the causality graph, optimizing parameters of the linear structure equation model by maximum likelihood estimation to obtain a structural causal model, the structural causal model being used for quantifying causal effect intensity among the variables in the causality graph;

15 computing the number of associated edges of each node in the causality graph, determining the nodes with the number of the associated edges exceeding a preset associated edge number threshold as decision nodes, tracing associated nodes level by level from the decision nodes along causality edges in the causality graph to obtain factor chains, computing a cumulative causal effect of the factor chains based on the structural causal model, and determining the factor chains with the cumulative causal effect exceeding a preset effect threshold as candidate decision basis chains;

20 in the structural causal model, changing the value of each node in the candidate decision basis chains and substituting into the structural causal model to obtain a corresponding downstream node predicted value after the value of the node is changed, determining the stability of the candidate decision basis chains based on a change amplitude of the predicted value, and determining the candidate decision basis chains with the stability meeting a preset condition as decision basis chains;

25 computing an influence degree of each node in the decision basis chains on a target variable based on the structural causal model, screening nodes with the influence degree exceeding a preset influence threshold as a feature set according to the influence degree, taking the influence degree of each node in the feature set as a feature importance score, and combining the feature importance score with the class discrimination of the classified and labelled data to obtain a feature optimization parameter;

30 performing weighted adjustment on the value of the corresponding feature in the initial structured data based on the feature optimization parameter to obtain optimized data, the feature optimization parameter being used for determining a weight coefficient of feature adjustment; and

35 computing the feature distribution difference between the optimized data and the

initial structured data to obtain an optimization effect index; inputting the optimized data into the structural causal model to verify causal relationship constraint so as to obtain a verification result, performing iterative optimization on the optimized data according to the optimization effect index and the verification result, and when the optimization index reaches a preset target and the verification result meets causal consistency, determining the current optimized data as the final structured data.

8. A deep learning-based automatic processing system for unstructured data, applied to implement the method according to any one of claims 1 to 7, comprising:

a first unit, configured to acquire unstructured data by an intelligent data collector; construct a data feature fingerprint library; perform classification and labelling on the unstructured data based on the data feature fingerprint library to obtain classified and labelled data; construct a data quality assessment model by an incremental learning algorithm, the data quality assessment model scoring the classified and labelled data from three dimensions of data integrity, timeliness and value density to obtain a scoring result; construct a priority processing queue according to the scoring result, and perform data cleaning and standardized processing on the classified and labelled data in the priority processing queue to obtain a standardized training data set;

a second unit, configured to perform deep learning processing on the standardized training data set by a multi-level cascade network structure, the multi-level cascade network structure identifying target features in the standardized training data set through a self-attention mechanism according to class information of the classified and labelled data; construct a feature association graph based on the target features; construct a bidirectional inference channel on the feature association graph, the bidirectional inference channel comprising a forward channel and a reverse channel, the forward channel extracting numerical distribution features, time-order features and attribute features in the standardized training data set by a convolutional neural network so as to obtain dominant feature data, and the reverse channel extracting data item association rules, data change trends and data interaction modes in the standardized training data set through a recurrent neural network so as to obtain recessive feature data; and perform feature combination on the dominant feature data and the recessive feature data by a dynamic weight allocation strategy to obtain a multi-dimensional feature tensor; and

a third unit, configured to perform rule matching on the multi-dimensional feature tensor based on a preset domain knowledge base to obtain an inference rule set; perform hierarchical processing on the multi-dimensional feature tensor by the inference rule set so as to obtain initial structured data; analyze the initial structured data by a causal inference

method to obtain a decision basis chain; generate optimization parameters based on the decision basis chain; and perform optimization processing on the initial structured data according to the optimization parameters and the class information of the classified and labelled data to obtain final structured data.

5 9. An electronic device, comprising:

 a processor; and

 a memory configured to store processor executable instructions;

 the processor is configured to call instructions stored in the memory to execute the method according to any one of claims 1 to 7.

10 10. A computer-readable storage medium having computer program instructions stored therein, wherein when executed by a processor, the computer program instructions execute the method according to any one of claims 1 to 7.