

# **TRANSFORMER-BASED MODELING AND ANALYSIS METHOD FOR STRUCTURED DATA**

## **TECHNICAL FIELD**

5        [0001] The present invention relates to the technical field of data analysis, and specifically relates to a Transformer-based modeling and analysis method for structured data.

## **BACKGROUND**

10        [0002] With the advent of the era of big data, the analysis and modeling of structured data have become an important research direction in the field of artificial intelligence. Traditional structured data analysis methods mainly rely on feature engineering and shallow machine learning models, such as decision trees and logistic regression. In recent years, a Transformer architecture has made groundbreaking  
15        progress in the field of natural language processing, and its self-attention mechanism can effectively capture long-range dependence relationship in sequential data. However, when applying Transformers to structured data analysis, it is needed to handle key problems such as the heterogeneous handling of numerical features and categorical features, feature interaction modeling, and model interpretability.

20        [0003] However, existing technologies still face problems in the field of structured data modeling and analysis, the traditional methods struggle to effectively handle the heterogeneity of the numerical features and the categorical features, and feature encoding methods are too simplistic to fully utilize association information among the features. Existing feature extraction methods mainly focus on local feature  
25        interactions and lack the capability to model global feature dependence relationship, resulting in limited feature representation capability. The feature selection process often employs a static scoring mechanism, ignoring the dynamic relevancy among the features and lacking interpretability support. The model prediction results lack fine-grained explanation mechanisms, making it difficult to quantify the contribution  
30        of each feature to the prediction results.

      [0004] In conclusion, there is an urgent need for a Transformer-based modeling and analysis method for structured data. A dual-stream feature encoder is designed to adaptively encode the numerical features and the categorical features separately. A hierarchical attention mechanism and a feature affinity graph are utilized to implement

multi-scale feature interaction modeling. An information bottleneck feature filter and a dual-branch prediction network are constructed to provide an interpretable feature importance assessment mechanism while ensuring model performance. Finally, the feature contribution is quantitatively analyzed through a structured gradient method, thus providing fine-grained interpretability for the model prediction results. The present invention not only improves the modeling effect of the structured data but also enhances the interpretability and reliability of the model, solving the problems in the related technology.

## 10 SUMMARY OF THE INVENTION

[0005] The embodiment of the present invention provides a Transformer-based modeling and analysis method for structured data, which can solve the problems in related technology.

[0006] In a first aspect,

15 [0007] the Transformer-based modeling and analysis method for structured data provided by the embodiment of the present invention includes:

[0008] receiving structured original data, dividing the structured original data into a numerical feature and a categorical feature, executing piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and executing adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; performing adaptive weight allocation by a learnable gating mechanism, and fusing the normalized main feature with the continuous value mapping feature to obtain a fusion feature; performing dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set;

25 [0009] inputting the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity matrix of sample features, constructing a feature affinity graph on the basis of the similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extracting a local feature representation by a causal attention mechanism, and integrating a plurality of local feature representations through a cross-sub-graph attention fusion mechanism to obtain a global feature representation; receiving the global feature representation by the

self-calibration module, and performing feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation; and

[0010] constructing an information bottleneck feature filter, and filtering the deep feature representation by maximizing task related information and minimizing  
 5 redundant information to obtain a key feature subset; inputting the key feature subset into a dual-branch network containing a main prediction branch and an auxiliary reconstruction branch, the main prediction branch processing the key feature subset based on the attention mechanism to obtain a feature weight vector, and generating a prediction result according to the feature weight vector, and the auxiliary  
 10 reconstruction branch providing a regularization constraint by reconstructing the key feature subset; and according to the feature weight vector and the prediction result, computing the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result.

[0011] In an optional embodiment,

15 [0012] the receiving structured original data, dividing the structured original data into a numerical feature and a categorical feature, executing piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and executing adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; performing adaptive weight  
 20 allocation by a learnable gating mechanism, and fusing the normalized main feature with the continuous value mapping feature to obtain a fusion feature; performing dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set includes:

[0013] receiving the structured original data, and dividing the structured original  
 25 data into the numerical feature and the categorical feature; inputting the numerical feature and the categorical feature into the dual-stream feature encoder for feature processing, the dual-stream feature encoder including a feature main stream and a feature mapping stream;

[0014] in the feature main stream, counting the data distribution of the numerical  
 30 feature to obtain data density, determining a location of an optimal segmentation point based on the data density, dividing the numerical feature into a plurality of segmentation intervals according to the optimal segmentation point, constructing a linear transformation parameter matrix in each segmentation interval, computing a feature difference value between adjacent segmentation intervals, determining an

optimized linear transformation parameter matrix of each segmentation interval by minimizing the feature difference value, and executing linear transformation on the numerical feature of each segmentation interval by the optimized linear transformation parameter matrix, so as to obtain the normalized main feature;

5 [0015] in the feature mapping stream, mapping each category value in the categorical feature into a corresponding quantum ground state, allocating an initial amplitude parameter and a phase parameter to each quantum ground state, constructing a quantum gate sequence containing a rotating gate and a control gate, sequentially acting the quantum gate sequence on the quantum ground states, performing quantum  
10 state transformation to obtain a transformed quantum state, computing an entropy value of the transformed quantum state, and projecting the transformed quantum state to a continuous space based on the entropy value to obtain the continuous value mapping feature;

[0016] computing an association coefficient between the normalized main feature  
15 and the continuous value mapping feature in each dimension respectively, combining the association coefficients to construct the feature association coefficient matrix, training a preset attention network based on the feature association coefficient matrix to obtain an importance vector of the feature dimension, and performing weighted combination on the normalized main feature and the continuous value mapping feature  
20 through the importance vector to obtain the fusion feature; and

[0017] computing the contribution of each dimension of the fusion feature to a target task, generating the feature importance score of each dimension based on the contribution, constructing a sampling probability distribution function according to the feature importance score, and performing importance sampling on training samples by  
25 the sampling probability distribution function to generate the enhanced training sample set.

[0018] In an optional embodiment,

[0019] the counting the data distribution of the numerical feature to obtain data density, determining a location of an optimal segmentation point based on the data  
30 density includes:

[0020] computing the width of a basic interval based on a interquartile range of the numerical feature and the number of samples, dividing the value range of the numerical feature into a plurality of basic statistical intervals based on the width of the basic interval, and counting the number of samples in each basic statistical interval to

obtain an interval frequency sequence;

[0021] counting a frequency ratio of adjacent basic statistical intervals in the interval frequency sequence to obtain the interval similarity, and merging the adjacent basic statistical intervals with the interval similarity greater than a preset merging  
5 threshold into a new statistical interval to obtain a merged frequency distribution sequence;

[0022] computing accumulated frequency distribution according to the merged frequency distribution sequence, and generating a data density function based on the accumulated frequency distribution; selecting candidate segmentation locations from  
10 the merged frequency distribution sequence, computing the number of samples and the numerical mean value of the intervals on two sides of each candidate segmentation location, and computing an interval variance value of the candidate segmentation locations by means of the number of samples and the numerical mean value;

[0023] selecting a location with the maximum interval variance value in the  
15 current intervals to be segmented to serve as a segmentation point to be determined, computing a ratio of the interval variance value of the segmentation point to be determined to the overall variance of the numerical feature, and determining a variance contribution ratio of the segmentation point to be determined; and

[0024] computing local distribution skewness of the intervals on the two sides of  
20 the segmentation point to be determined, when the difference value of the local distribution skewness is greater than a preset skewness threshold and the variance contribution ratio is greater than a preset segmentation threshold, determining the segmentation point to be determined to be the optimal segmentation point, and repeating the steps to determine the locations of all the optimal segmentation points.

25 [0025] In an optional embodiment,

[0026] the inputting the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity matrix of sample features, constructing a feature affinity graph on the basis of the similarity matrix, and dividing  
30 the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extracting a local feature representation by a causal attention mechanism, and integrating a plurality of local feature representations through a cross-sub-graph attention fusion mechanism to obtain a global feature representation; receiving the global feature representation by the

self-calibration module, and performing feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation includes:

[0027] inputting the enhanced training sample set into a convolutional neural network to obtain an initial feature map, computing a dot product of any two sample  
 5 features in the initial feature map, and dividing the dot product by a module length product of a corresponding feature vector to obtain the feature similarity matrix; constructing the feature affinity graph based on the feature similarity matrix, a vertex of the feature affinity graph corresponding to the sample feature, an edge weight of the feature affinity graph corresponding to a similarity value in the feature similarity  
 10 matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs by a spectral clustering algorithm;

[0028] in each feature sub-graph, dividing the feature sequence of the feature sub-graph into a query feature sequence and a key-value pair feature sequence, constructing a time-order causal mask matrix, and shielding an interaction result of the  
 15 query feature sequence and the key-value pair feature sequence by the time-order causal mask matrix to obtain an initial attention weight only containing historical feature information; computing a spatial distance matrix of the sample features in each feature sub-graph, computing a spatial distribution entropy of the feature sub-graph based on the spatial distance matrix, and performing adaptive adjustment on the initial  
 20 attention weight according to the spatial distribution entropy to obtain the adaptive attention weight; weighting the key-value pair feature sequence based on the adaptive attention weight, and superposing the location encoding information to obtain the local feature representation of each feature sub-graph;

[0029] based on a pre-constructed multilayer perceptron network, inputting a  
 25 splicing vector of the local feature representation of adjacent feature sub-graphs, performing forward transmission through the multilayer perceptron network to obtain an association weight between the feature sub-graphs, and performing weighted summation on the local feature representation based on the association weight to obtain the global feature representation;

[0030] computing a mean value and a standard deviation of the global feature  
 30 representation in the feature dimension; based on a pre-constructed dynamic normalization network, inputting the global feature representation into the dynamic normalization network, and performing forward transmission to obtain a feature transformation coefficient and a feature offset; subtracting the mean value from the

global feature representation and dividing the result by the standard deviation to obtain a standardized feature, multiplying the standardized feature by the feature transformation coefficient, and adding the product to the feature offset to obtain the transformation feature, adding the transformation feature and the global feature representation to obtain a self-calibration feature; and

[0031] performing layer normalization processing on the self-calibration feature, and performing nonlinear transformation through a ReLU activation function to obtain the deep feature representation.

[0032] In an optional embodiment,

[0033] the dividing the feature sequence of the feature sub-graph into a query feature sequence and a key-value pair feature sequence, constructing a time-order causal mask matrix, and shielding an interaction result of the query feature sequence and the key-value pair feature sequence by the time-order causal mask matrix to obtain an initial attention weight only containing historical feature information includes:

[0034] based on the similarity of adjacent time step features in the feature sub-graphs, grouping the feature sequence of the feature sub-graphs to obtain the query feature sequence and the key-value pair feature sequence, classifying a time step feature with the similarity of the adjacent time step features greater than a dynamic threshold determined based on the global statistical character of the feature sequence into the query feature sequence, classifying the rest of the features into the key-value pair feature sequence, so as to obtain the feature classification result of the query feature sequence and the key-value pair feature sequence;

[0035] based on the feature classification result, computing the time-order association intensity of the query feature sequence and the key-value pair feature sequence through the recurrent neural network, a hidden state of the recurrent neural network containing long short term dependence information of the query feature sequence and the key-value pair feature sequence, so as to obtain the time-order association intensity of the query feature sequence and the time-order association intensity of the key-value pair feature sequence;

[0036] constructing a first mask matrix based on the time-order location relationship of the feature sequence to serve as a time-order causal mask matrix, transforming the time-order association intensity of the query feature sequence and the time-order association intensity of the key-value pair feature sequence into soft mask weights to obtain a second mask matrix; and

[0037] performing matrix multiplication operation on the query feature sequence and the key-value pair feature sequence to obtain an original attention score; performing addition operation on the original attention score and the first mask matrix to obtain a first mask result, performing element-level multiplication operation on the first mask result and the second mask matrix to obtain an attention score obtained after dual mask; and performing normalization processing on the attention score obtained after dual mask to obtain the initial attention weight only containing the historical feature information.

[0038] In an optional embodiment,

[0039] the constructing an information bottleneck feature filter, and filtering the deep feature representation by maximizing task related information and minimizing redundant information to obtain a key feature subset includes:

[0040] constructing the information bottleneck encoder, the information bottleneck encoder including a three-layer perceptron structure, a normalization layer being connected behind each layer of perceptron structure, inputting the deep feature representation into the information bottleneck encoder, and generating probability distribution parameters of a compression feature space at an output layer of the information bottleneck encoder; performing code mapping on the deep feature representation based on the probability distribution parameters to obtain a compression feature representation;

[0041] constructing a mutual information assessment module, the mutual information assessment module containing the three-layer perceptron structures, inputting a splicing vector of the compression feature representation and a tag of the target task into the mutual information assessment module, and outputting mutual information amount between the compression feature representation and the target task;

[0042] computing an autocorrelation matrix of the compression feature representation, each element in the autocorrelation matrix being obtained by dividing a product that is obtained by subtracting a corresponding mean value from a corresponding feature vector by a product of the corresponding standard deviation; computing feature redundancy based on a difference value between the autocorrelation matrix and a unit matrix;

[0043] constructing a feature selection network, the feature selection network including a first fully-connected layer and a second fully-connected layer, inputting the

compression feature representation into the first fully-connected layer, inputting the output of the first fully-connected layer into the second fully-connected layer after processing through the activation function, processing the output of the second fully-connected layer through a gating function to obtain a gating value, and  
 5 performing element-by-element multiplication on the gating value and the compression feature representation to obtain an initial filter feature;

[0044] adjusting the output of the gating function based on temperature parameters, the temperature parameters decreasing progressively along with the number of training rounds based on a preset decreasing amount, and performing  
 10 element-by-element multiplication on the adjusted gating value and the compression feature representation to obtain an adjusted filter feature;

[0045] performing linear combination of an opposite number of the mutual information amount, the feature redundancy and a norm of the adjusted filter feature to construct a target function;

15 [0046] optimizing the target function by iteration to obtain an optimization parameter of the information bottleneck encoder and an optimization parameter of the feature selection network, so as to obtain an optimized information bottleneck encoder and an optimized feature selection network; and

[0047] inputting the deep feature representation into the optimized information  
 20 bottleneck encoder to obtain the optimized compression feature representation, and inputting the optimized compression feature representation into the optimized feature selection network to obtain the key feature subset.

[0048] In an optional embodiment,

[0049] the according to the feature weight vector and the prediction result,  
 25 computing the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result includes:

[0050] based on the feature weight vector and the prediction result, computing the association coefficient between each pair of features in the feature weight vector, so as to generate the feature association matrix; constructing a feature dependency graph  
 30 based on the feature association matrix, treating each feature in the key feature subset as a node in the feature dependency graph, and treating the association coefficient in the feature association matrix serves as a weight of an edge between the corresponding nodes;

[0051] constructing a graph Laplacian matrix according to the feature dependency

graph, treating the sum of the edge weight of each node and the edge weight of the corresponding connected node as a diagonal element value of the graph Laplacian matrix, and treating a negative value of the edge weights between the nodes as a non-diagonal element value of the graph Laplacian matrix;

5        [0052] computing a Jacobian matrix of the prediction result about the key feature subset to obtain an initial gradient; treating the graph Laplacian matrix as a structured regularization item, and performing weighted combination on the structured regularization item and the initial gradient to obtain the structured gradient under the feature dependence relationship;

10       [0053] computing a structured gradient amplitude of each feature in the key feature subset, an interaction gradient value of each feature and other features, and centrality measurement of each feature in the feature dependency graph; and

      [0054] performing weighted combination on the structured gradient amplitudes, the interaction gradient values and the centrality measurements to obtain the feature contribution; performing normalization processing on the feature contribution to obtain a standardized contribution of each feature in the key feature subset to the prediction result; and performing importance sorting on the features in the key feature subset based on the standardized contribution to generate the analysis result.

      [0055] In a second aspect,

20       [0056] a Transformer-based modeling and analysis system for structured data provided by an embodiment of the present invention includes:

      [0057] a first unit, configured to receive structured original data, divide the structured original data into a numerical feature and a categorical feature, execute piecewise linear transformation on the numerical feature by a dual-stream feature  
25       encoder to obtain a normalized main feature, and execute adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; perform adaptive weight allocation by a learnable gating mechanism, and fuse the normalized main feature with the continuous value mapping feature to obtain a fusion feature; and perform dynamic sampling on the basis of an importance score of the fusion feature to  
30       generate an enhanced training sample set;

      [0058] a second unit, configured to input the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity matrix of sample features, constructing a feature affinity graph on the basis of the

similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extract local feature representation by a causal attention mechanism, and integrate a plurality of local feature representation through a cross-sub-graph attention fusion mechanism to  
5 obtain a global feature representation; and receive the global feature representation by the self-calibration module, and perform feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation; and

[0059] a third unit, configured to construct an information bottleneck feature filter, and filter the deep feature representation by maximizing task related information and  
10 minimizing redundant information to obtain a key feature subset; input the key feature subset into a dual-branch network containing a main prediction branch and an auxiliary reconstruction branch, the main prediction branch processing the key feature subset based on the attention mechanism to obtain a feature weight vector, and generating a prediction result according to the feature weight vector, and the auxiliary  
15 reconstruction branch providing a regularization constraint by reconstructing the key feature subset; and according to the feature weight vector and the prediction result, compute the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result.

[0060] In a third aspect,

20 [0061] an electronic device provided by the embodiment of the present invention includes:

[0062] a processor; and

[0063] a memory configured to store processor executable instructions; and

[0064] the processor is configured to call instructions stored in the memory to  
25 execute the abovementioned method.

[0065] In a fourth aspect,

[0066] 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  
30 method.

[0067] In the embodiment of the present invention, the dual-stream feature encoder is configured to process the numerical feature and the categorical feature respectively, and the learnable gating mechanism is utilized to perform adaptive fusion, so that the processing capability for different types of data is effectively improved;

dynamic sampling is performed based on the importance score of the fusion feature, the enhanced training sample set is generated, and thus the generalization capability and robustness of the model are remarkably improved; the feature extraction network containing the hierarchical attention module and the self-calibration module is utilized, the complex relationship between the features is captured through the feature affinity graph and the graph segmentation algorithm, and the causal attention mechanism and cross-sub-graph attention fusion are utilized to extract the global feature representation; the self-calibration module further optimizes the feature representation by means of residual connection and dynamic normalization, so that the accuracy of feature extraction and the expression capability of the model are effectively improved; the information bottleneck feature filter is added, thus the redundant information is minimized while the task related information is reserved, and therefore, the feature selection efficiency is improved; due to the design of the dual-branch network, the prediction accuracy is guaranteed, regularization constraints are provided through the auxiliary reconstruction branches, thereby enhancing the stability of the model; and the structured gradient method is adopted to compute the feature contribution, which provides interpretability for model decision, and facilitates a user to understand and trust the prediction result of the model.

## **BRIEF DESCRIPTION OF THE DRAWINGS**

[0068] FIG. 1 is a flowchart of a Transformer-based modeling and analysis method for structured data according to an embodiment of the present invention; and

[0069] FIG. 2 is a schematic structural diagram of a Transformer-based modeling and analysis system for structured data according to an embodiment of the present invention.

## **DETAILED DESCRIPTION**

[0070] 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.

[0071] 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.

[0072] FIG. 1 is a flowchart of a Transformer-based modeling and analysis method for structured data according to an embodiment of the present invention, as shown in FIG. 1, the method includes:

[0073] S101: receiving structured original data, dividing the structured original data into a numerical feature and a categorical feature, executing piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and executing adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; performing adaptive weight allocation by a learnable gating mechanism, and fusing the normalized main feature with the continuous value mapping feature to obtain a fusion feature; performing dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set;

[0074] in this embodiment, the numerical feature and the categorical feature are processed respectively by the dual-stream feature encoder, and suitable encoding modes can be adopted for different feature types, thereby improving the accuracy of feature expression; adaptive weights are allocated to different feature types through the learnable gating mechanism, thus effective fusion of the numerical feature and the categorical feature is achieved, multi-source information is fully utilized, and the expressive force of the model is improved; dynamic sample optimization is performed: dynamic sampling is performed based on the importance score of the fusion feature, which is conducive to balancing sample distribution, and highlighting the key samples, thereby generating a more representative enhanced training sample set, and improving the training effect and generalization capability of the model; the whole method is to optimize in the data preprocessing and feature expression stages, thus more accurate and efficient feature input is provided for downstream tasks, and the learning efficiency and performance representation of the model are further improved.

[0075] S102: inputting the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity matrix of sample features,

constructing a feature affinity graph on the basis of the similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extracting a local feature representation by a causal attention mechanism, and integrating a plurality of local feature representations through a cross-sub-graph attention fusion mechanism to obtain a global feature representation; receiving the global feature representation by the self-calibration module, and performing feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation;

[0076] in this embodiment, the feature affinity graph is constructed by the hierarchical attention module and is divided into the plurality of feature sub-graphs, thus the structured association among the features can be captured, and the modeling capability for the internal relationship of the samples can be improved; the local feature representation is extracted through the causal attention mechanism, the local features are integrated through the cross-sub-graph attention fusion mechanism, and local details and a global mode can be effectively combined, thereby improving the comprehensiveness of feature expression; the calibration module implements nonlinear transformation and adaptive adjustment for the global feature representation by means of residual connection and dynamic normalization, which is conducive to enhancing the key features and inhibiting redundant information, as well as improving the representation capability for deep features; the whole network architecture is combined with hierarchical attention and the self-calibration module, therefore, the multi-dimensional features of complex samples can be extracted more efficiently, and deep feature representation with high recognition degree can be provided for downstream tasks; the feature extraction network enhances the modeling capability for the data relationship by means of graph segmentation, attention mechanism and dynamic calibration, and thus the modeling capability for the complex data relationship is enhanced, and the adaptability and generalization capability of the model to diversified input are improved.

[0077] S103: constructing an information bottleneck feature filter, and filtering the deep feature representation by maximizing task related information and minimizing redundant information to obtain a key feature subset; inputting the key feature subset into a dual-branch network containing a main prediction branch and an auxiliary reconstruction branch, the main prediction branch processing the key feature subset based on the attention mechanism to obtain a feature weight vector, and generating a

prediction result according to the feature weight vector, and the auxiliary reconstruction branch providing a regularization constraint by reconstructing the key feature subset; and according to the feature weight vector and the prediction result, computing the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result.

[0078] Firstly, the information bottleneck feature filter is constructed for analyzing and filtering the deep feature representation. The mutual information value between each feature dimension and the prediction target is computed, the higher the mutual information value is, the stronger the relevancy between the dimension feature and the prediction task is. Moreover, the redundant information among the feature dimensions is computed for discovering and removing redundant features with overlapped information. By the iterative optimization process, the key feature subset can be filtered from the complete deep feature representation under the constraints of maximizing the task related information and minimizing the feature redundant information.

[0079] The key feature subset obtained by filtering is inputted into the dual-branch network to be processed. In a main prediction branch, a multi-head attention mechanism is utilized to perform weighting on the features. The multi-head attention layer contains a plurality of parallel attention computation units, and each attention head independently computes the feature weight. In this way, the association mode between the features can be captured by the model from different angles. The outputs of all attention heads are merged to obtain the comprehensive feature weight vector. On the basis of the weight vector, weighted combination is performed on the key features, and the prediction result is generated by the fully-connected layer.

[0080] In the auxiliary reconstruction branch, a symmetrical encoder-decoder structure is constructed. The key features are mapped to a low-dimensional hidden space by the encoder using the multilayer fully-connected network, and the hidden space representation is reconstructed back to an original feature space by the decoder. This reconstruction task is treated as an auxiliary supervision signal, and the learning process of the model is retrained by a minimum reconstruction error. The reconstruction loss and the prediction loss jointly form the optimization target of the model, so that the feature representation obtained by learning is favorable for the prediction task, and the structure information of original data is also kept.

[0081] Finally, the prediction result is interpreted and analyzed by the structured

gradient method. The contribution of each input feature to the prediction result is computed in combination with the feature weight vector obtained by the main branch. The feature contribution reflects the importance degree of different features in the decision process and helps to understand the prediction basis of the model. And the interpretable result report is generated based on contribution analysis, which helps the user to understand the key influence factor behind the prediction.

[0082] In this embodiment, the task related information is maximized and the redundant information is minimized by the information bottleneck feature filter, the key feature subset is extracted from the deep feature representation, and thus the efficiency and precision of feature selection are remarkably improved; the main prediction branch provides the weight to the key feature subset based on the attention mechanism, and a more accurate prediction result is generated by focusing on the features having great influences on the prediction task; the auxiliary reconstruction branch adds the regularization constraints by reconstructing the key feature subset, thus preventing the model from overfitting, as well as ensuring the integrity and expression capability of the key feature subset; the contribution of each key feature in the key feature subset to the prediction result is quantified by the structured gradient method, thus the support is provided for model interpreting performance and interpretability, and the decision basis of the model is helped to be understood; and by filtering and optimization, the features highly related to the task target are reserved and enhanced, and therefore the task adaptation capability and practical application value of the model are improved.

[0083] In an optional embodiment, the receiving structured original data, dividing the structured original data into a numerical feature and a categorical feature, executing piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and executing adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; performing adaptive weight allocation by a learnable gating mechanism, and fusing the normalized main feature with the continuous value mapping feature to obtain a fusion feature; performing dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set includes:

[0084] receiving the structured original data, and dividing the structured original data into the numerical feature and the categorical feature; inputting the numerical feature and the categorical feature into the dual-stream feature encoder for feature

processing, the dual-stream feature encoder including a feature main stream and a feature mapping stream;

[0085] in the feature main stream, counting the data distribution of the numerical feature to obtain data density, determining a location of an optimal segmentation point  
 5 based on the data density, dividing the numerical feature into a plurality of segmentation intervals according to the optimal segmentation point, constructing a linear transformation parameter matrix in each segmentation interval, computing a feature difference value between adjacent segmentation intervals, determining an optimized linear transformation parameter matrix of each segmentation interval by  
 10 minimizing the feature difference value, and executing linear transformation on the numerical feature of each segmentation interval by the optimized linear transformation parameter matrix, so as to obtain the normalized main feature;

[0086] in the feature mapping stream, mapping each category value in the categorical feature into a corresponding quantum ground state, allocating an initial  
 15 amplitude parameter and a phase parameter to each quantum ground state, constructing a quantum gate sequence containing a rotating gate and a control gate, sequentially acting the quantum gate sequence on the quantum ground states, performing quantum state transformation to obtain a transformed quantum state, computing an entropy value of the transformed quantum state, and projecting the transformed quantum state  
 20 to a continuous space based on the entropy value to obtain the continuous value mapping feature;

[0087] computing an association coefficient between the normalized main feature and the continuous value mapping feature in each dimension respectively, combining the association coefficients to construct the feature association coefficient matrix,  
 25 training a preset attention network based on the feature association coefficient matrix to obtain an importance vector of the feature dimension, and performing weighted combination on the normalized main feature and the continuous value mapping feature through the importance vector to obtain the fusion feature; and

[0088] computing the contribution of each dimension of the fusion feature to a  
 30 target task, generating the feature importance score of each dimension based on the contribution, constructing a sampling probability distribution function according to the feature importance score, and performing importance sampling on training samples by the sampling probability distribution function to generate the enhanced training sample set.

[0089] In a specific embodiment, firstly, the structured raw data is received, and the features are divided according to the data type. Taking a user purchase intention prediction task of an e-commerce platform as an example, the raw data contains basic user information, behavioral data, and transaction records. The numerical features  
 5 include: age (18-80), number of visits in the past 30 days (0-120), total consumption in the past 90 days (0-50,000), days since last purchase (0-365), number of favorited items (0-100), number of items in the shopping cart (0-50), etc.; and the categorical features include: gender (male/female), membership level (regular/silver/gold/diamond), commonly used device type (PC/Android/iOS),  
 10 preferred shopping category (clothing/digital/beauty/books, etc.), delivery region, and so on.

[0090] In the feature main stream, firstly, distribution statistics and segmentation processing are performed on the numerical features: taking the user age feature as an example, the data distribution density of the age is obtained by counting based on  
 15 500,000 user samples. The width of the basic statistics interval is set to be 2 years old, and it is preliminarily divided into 31 basic intervals (18-20, 20-22...78-80). The sample frequency ratio of adjacent intervals is computed, and interval merging is performed when the ratio is greater than 0.85. For example, the frequencies of the three intervals 20-22, 22-24 and 24-26 are 15,000, 14,500, and 14,800 respectively, the  
 20 ratios of two adjacent groups are 0.97 and 0.98 respectively, and the intervals are merged into the interval 20-26 when merging conditions are met. Finally seven segmented intervals 18-20, 20-26, 26-35, 35-45, 45-55, 55-65, and 65-80 are obtained after merging.

[0091] The linear transformation parameter matrix is constructed for each  
 25 segmented interval: taking the interval of 20-26 years old as an example, a linear transformation parameter matrix of  $2 \times 2$  is constructed, and the initial value is the unit matrix. The feature difference value relative to the adjacent interval 26-35 is computed, and the difference value is enabled to be less than a threshold of 0.01 by iterative optimization. Finally, the obtained transformation matrix is applied to an age value in  
 30 this interval, thereby realizing normalized transformation of the features. For example, a normalized value of 0.37 is obtained after transformation of the original age of 24 years old. Other numerical features are processed by the same method, thereby obtaining the complete normalized main feature set.

[0092] In the feature mapping stream, quantum transcoding is performed on the

categorical feature: taking the member level feature as an example, firstly, four levels are mapped into corresponding quantum ground states: the regular member is mapped into  $|00\rangle$ , the silver member is mapped into  $|01\rangle$ , the gold member is mapped into  $|10\rangle$ , and the diamond member is mapped into  $|11\rangle$ . The initial parameters are allocated for each ground state: all the amplitudes are set to be 0.5, and a phase angle is set to be 0. A sequence containing 8 quantum gates is constructed, namely: an X gate, a CNOT gate, a Z gate, an X gate, a CNOT gate, a Z gate, an X gate, and a CNOT gate in sequence. This gate sequence is sequentially acted on the ground states to obtain the transformed quantum state. The entropy value of the quantum state is computed, the entropy value of the regular member is 0.25, the entropy value of the silver member is 0.45, the entropy value of the gold member is 0.65, and the entropy value of the diamond member is 0.85. The quantum state is projected to a two-dimensional continuous space based on the entropy value to obtain the continuous value representation of the member level, for example, the regular member is mapped into (0.2, 0.3), and the silver member is mapped into (0.4, 0.5).

[0093] The feature relevancy analysis and fusion are performed: the association coefficient between every two of all the features is computed. For example, the association coefficient between the age and the total consumption is 0.35, and the association coefficient between the member level and the number of visit is 0.62. These association coefficients are organized into the relevancy matrix, and the matrix dimension is the total number of features x the total number of features. An attention network containing a three-layer structure is trained based on the matrix, the number of hidden layer units of the network is 128, 64, and 32 respectively, and the ReLU activation function is used. The importance vector of the feature dimension is obtained by network training, for example, the importance of the age feature is 0.28, the importance of the total consumption is 0.35, and the importance of the member level is 0.31. Weighted combination is performed on the normalized main feature and the continuous value mapping feature through the importance values to obtain the final fusion feature.

[0094] Finally, the sample enhancement processing is performed: the contribution of each dimension in the fusion feature to the purchase intention prediction is computed. Based on 500,000 training samples, it is computed to obtain the contribution value of each feature dimension: 0.25 of the age feature, 0.38 of the total consumption, 0.32 of the member level, 0.28 of the number of visit, and the like. These

contribution values are normalized to serve as the sampling probability, and the sampling probability distribution function is constructed. Importance sampling is performed on the original training sample set, 3 samples with the sampling probability exceeding 0.8 are copied, 2 samples with the probability between 0.5 and 0.8 are copied, and 1 sample with the probability between 0.3 and 0.5 is copied. In this way, the enhanced training sample set is generated, the final number of samples is 1.8 times that of the original data set, and about 900,000 sample records are generated.

[0095] In this embodiment, the numerical features and the categorical features are processed by the dual-stream feature encoder respectively, so that efficient encoding of different types of features is realized. The continuity and local features of the numerical features are kept by piecewise linear transformation, and the advantages of quantum computing are fully utilized in the adaptive quantum encoding, so that the expression capability of the categorical features is improved; feature fusion is performed through the learnable gating mechanism, and thus dynamic weight allocation among the features is realized. The importance degrees of different features are adaptively adjusted by learning the feature relevancy through the attention network, which improves the adaptability of the model to different feature combinations; and the quality of the training samples is effectively improved based on the dynamic sampling strategy of the feature importance. By computing the contribution of the features to the target task, reasonable sampling probability distribution is constructed, more representative training samples are generated, and therefore, the generalization capability and robustness of the model are improved.

[0096] In an optional embodiment, the counting the data distribution of the numerical feature to obtain data density, determining a location of an optimal segmentation point based on the data density includes:

[0097] computing the width of a basic interval based on a interquartile range of the numerical feature and the number of samples, dividing the value range of the numerical feature into a plurality of basic statistical intervals based on the width of the basic interval, and counting the number of samples in each basic statistical interval to obtain an interval frequency sequence;

[0098] counting a frequency ratio of adjacent basic statistical intervals in the interval frequency sequence to obtain the interval similarity, and merging the adjacent basic statistical intervals with the interval similarity greater than a preset merging threshold into a new statistical interval to obtain a merged frequency distribution

sequence;

[0099] computing accumulated frequency distribution according to the merged frequency distribution sequence, and generating a data density function based on the accumulated frequency distribution; selecting candidate segmentation locations from the merged frequency distribution sequence, computing the number of samples and the numerical mean value of the intervals on two sides of each candidate segmentation location, and computing an interval variance value of the candidate segmentation locations by means of the number of samples and the numerical mean value;

[0100] selecting a location with the maximum interval variance value in the current intervals to be segmented to serve as a segmentation point to be determined, computing a ratio of the interval variance value of the segmentation point to be determined to the overall variance of the numerical feature, and determining a variance contribution ratio of the segmentation point to be determined; and

[0101] computing local distribution skewness of the intervals on the two sides of the segmentation point to be determined, when the difference value of the local distribution skewness is greater than a preset skewness threshold and the variance contribution ratio is greater than a preset segmentation threshold, determining the segmentation point to be determined to be the optimal segmentation point, and repeating the steps to determine the locations of all the optimal segmentation points.

[0102] The local distribution skewness specifically refers to an index for measuring whether the data distribution in a certain interval is symmetrical or not, and it is used for describing the deviation direction and degree of the data distribution in the interval. If the data distribution is bilaterally symmetrical, the skewness will be zero; if the data distribution is deviated rightwards, it is indicated that there are fewer larger numerical values but the tail is lengthened, and the skewness is positive; if the data distribution is deviated leftwards, it is indicated that there are fewer smaller numerical values but the tail is longer, and the skewness is negative. The local distribution skewness focuses on the distribution characteristic in a certain interval instead of the overall distribution condition of the data, and therefore the distribution features of the data in different intervals can be analyzed more carefully.

[0103] In a specific embodiment, firstly, statistical analysis is performed on the data distribution of the numerical features. The overall value range is determined by computing the maximum value and the minimum value of the numerical features. The width of the basic statistical interval is set based on the interquartile range of the data

and the total number of samples. For example, for a data set with the number of samples of ten thousand, the interquartile range of the age feature is 20 years old, and the width of the basic interval can be set to be 5 years old. According to the width of this basic interval, the value range is divided into a plurality of basic statistical intervals, and the number of samples in each interval is counted, thus obtaining the initial interval frequency sequence.

[0104] Then, the data distribution similarity of the adjacent basic statistical intervals is analyzed. The frequency ratio of the adjacent intervals is computed, and the merging threshold is set to be 0.8. Taking the age feature as an example, if the number of samples in the interval from 20 years old to 25 years old is 1,000, and the number of samples in the interval from 25 years old to 30 years old is 900, the similarity of the two intervals is 0.9, which is greater than the merging threshold, and the two intervals can be merged into a new statistical interval. All the adjacent intervals are sequentially processed to obtain the frequency distribution sequence after merging.

[0105] Based on the frequency distribution sequence after merging, the cumulative frequency of each interval is computed, and the data density function is constructed. A location with obvious frequency change in the frequency distribution sequence is selected as the candidate segmentation location. For each candidate location, the number of samples and the numerical mean value of the intervals on the left and right sides of the candidate location are computed respectively. The interval variance value of this location is computed to represent the distinguishing effect of the location serving as the segmentation point.

[0106] In the interval to be segmented currently, the location with the maximum interval variance value is selected as the segmentation point to be determined. The ratio of the interval variance value corresponding to the segmentation point to be determined to the total feature variance is computed to obtain the variance contribution ratio. Moreover, the data distribution skewness of the intervals on the two sides of the segmentation point to be determined is computed. The skewness threshold is set to be 0.5, and the segmentation threshold is set to be 0.3. When the skewness difference value of the intervals on the two sides is greater than the skewness threshold and the variance contribution ratio is greater than the segmentation threshold, it is determined that the segmentation point to be determined is the optimal segmentation point.

[0107] The above analysis process is repeated until the segmentation effects of all the intervals do not meet the threshold requirements, thereby obtaining the locations of

all the optimal segmentation points. Taking the age feature as an example, the possible optimal segmentation points include 18 years old, 35 years old, 50 years old, and the like, and these segmentation points divide the sample data into a plurality of intervals having remarkable statistical feature differences.

5        [0108] In this embodiment, interval merging is performed by computing the frequency ratio of adjacent intervals, so that the interference of intervals with sparse data on the segmentation result is avoided, and the stability and reliability of the segmentation result are improved; the interval variance value and the variance contribution ratio are adopted as the segmentation point assessment indexes, and in  
10 combination with the difference constraints of the interval distribution skewness, the statistical significance of the segmentation point locations is ensured, and thus the segmentation result has good distinction degree; and the optimal segmentation points are adaptively determined based on the data distribution features, it is not needed to manually specify the number of segments, thus the influence of subjective factors is  
15 avoided, and the applicability of the segmentation method in different scenarios is improved.

      [0109] In an optional embodiment, the inputting the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity  
20 matrix of sample features, constructing a feature affinity graph on the basis of the similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extracting a local feature representation by a causal attention mechanism, and integrating a plurality of local feature representations through a cross-sub-graph  
25 attention fusion mechanism to obtain a global feature representation; receiving the global feature representation by the self-calibration module, and performing feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation includes:

      [0110] inputting the enhanced training sample set into a convolutional neural  
30 network to obtain an initial feature map, computing a dot product of any two sample features in the initial feature map, and dividing the dot product by a module length product of a corresponding feature vector to obtain the feature similarity matrix; constructing the feature affinity graph based on the feature similarity matrix, a vertex of the feature affinity graph corresponding to the sample feature, an edge weight of the

feature affinity graph corresponding to a similarity value in the feature similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs by a spectral clustering algorithm;

[0111] in each feature sub-graph, dividing the feature sequence of the feature sub-graph into a query feature sequence and a key-value pair feature sequence, constructing a time-order causal mask matrix, and shielding an interaction result of the query feature sequence and the key-value pair feature sequence by the time-order causal mask matrix to obtain an initial attention weight only containing historical feature information; computing a spatial distance matrix of the sample features in each feature sub-graph, computing a spatial distribution entropy of the feature sub-graph based on the spatial distance matrix, and performing adaptive adjustment on the initial attention weight according to the spatial distribution entropy to obtain the adaptive attention weight; weighting the key-value pair feature sequence based on the adaptive attention weight, and superposing the location encoding information to obtain the local feature representation of each feature sub-graph;

[0112] based on a pre-constructed multilayer perceptron network, inputting a splicing vector of the local feature representation of adjacent feature sub-graphs, performing forward transmission through the multilayer perceptron network to obtain an association weight between the feature sub-graphs, and performing weighted summation on the local feature representation based on the association weight to obtain the global feature representation;

[0113] computing a mean value and a standard deviation of the global feature representation in the feature dimension; based on a pre-constructed dynamic normalization network, inputting the global feature representation into the dynamic normalization network, and performing forward transmission to obtain a feature transformation coefficient and a feature offset; subtracting the mean value from the global feature representation and dividing the result by the standard deviation to obtain a standardized feature, multiplying the standardized feature by the feature transformation coefficient, and adding the feature offset to obtain the transformation feature, adding the transformation feature and the global feature representation to obtain a self-calibration feature; and

[0114] performing layer normalization processing on the self-calibration feature, and performing nonlinear transformation through a ReLU activation function to obtain the deep feature representation.

[0115] In a specific embodiment, firstly, data enhancement processing is performed on the inputted training sample set. Data enhancement includes operations of random cropping, horizontal overturning, brightness and contrast adjustment and the like, thus generating the enhanced training sample set. The number of enhanced samples is 4 times that of the original samples, and each original sample generates 3 new samples by different enhancement modes.

[0116] The enhanced training samples are inputted into the convolutional neural network for feature extraction. ResNet50 is adopted as the basic network for performing multi-scale feature extraction. The size of the feature graph outputted at the last convolutional layer is  $14 \times 14$ , and the number of channels is 2,048. For each sample, the feature graph is flattened into a one-dimensional vector, and thus the initial feature map is obtained.

[0117] When computing the feature similarity matrix, dot product operation is performed on the feature vectors of any two samples, and the result is divided by respective module length product for normalization. For example, for two 2048-dimensional feature vectors, their cosine similarity is computed, and the obtained similarity value range is between -1 and 1. The feature affinity graph is constructed based on the similarity matrix, vertexes in the graph represent the sample features, and the weights of the edges are the corresponding similarity values.

[0118] The feature affinity graph is divided by the spectral clustering algorithm. Firstly, the Laplacian matrix of the graph is computed, and the feature vectors corresponding to  $k$  minimum feature values are selected,  $k$  is the number of preset sub-graphs and is usually set to be between 4 and 8. K-means clustering is performed on the feature vectors, and thus the samples are allocated into different feature sub-graphs.

[0119] The causal attention mechanism is implemented in each feature sub-graph. The feature sequence is equally divided into the query sequence and the key-value pair sequence. An upper triangular causal mask matrix is constructed, and it is ensured that only historical information can be visited at the current moment. The Euclidean distances between the features are computed to construct the spatial distance matrix, and the normalized spatial distribution entropy is computed. According to the entropy value, the attention weights are adaptively adjusted, the larger the entropy value is, the more dispersive the feature distribution is, and the smaller the corresponding attention weight is.

[0120] The three-layer multilayer perceptron network is constructed for attention fusion among the feature sub-graphs. The dimensions of the hidden layer are 1024, 512 and 256 respectively. The local feature splicing vectors of adjacent sub-graphs are inputted, and the association weights thereof are outputted. Weighted summation is performed on all local features according to the association weights, so as to obtain the global feature representation.

[0121] In the self-calibration module, firstly, the statistical magnitude of the global features in the feature dimensions is computed. The dual-layer dynamic normalization network is constructed, and the dimension of the hidden layer is 1024. The feature transformation coefficient and the offset are outputted by the network for performing adaptive adjustment on the standardized features. The adjusted features are added to the original features to realize residual connection.

[0122] Finally, layer normalization is performed on the self-calibration features, and the final deep feature representation is obtained through the ReLU activation function. The feature dimensions are kept at 2048 for subsequent task.

[0123] By way of example, taking the user purchase intention prediction task of the e-commerce platform as an example, firstly, 900,000 enhanced training samples are inputted into the convolutional neural network. The convolutional neural network contains 3 convolutional layers, the sizes of convolutional kernels are 3 x 3, 3 x 3, and 5 x 5 respectively, and the numbers of toutput channels are 64, 128, and 256 respectively. After processing by the convolutional network, the initial feature map with the dimension of 256 is obtained. The dot product similarity of any two sample features is computed, and normalization processing is carried out to obtain a feature similarity matrix with the size of 900,000 x 900,000.

[0124] The feature affinity graph is constructed based on the feature similarity matrix. Each sample feature is treated as a vertex of the graph, and the similarity value between the samples is treated as the weight of the edge. For example, if the similarity between the feature vector of a user A and the feature vector of a user B is 0.85, the weight of the edge between the two vertexes in the affinity graph will be 0.85. The feature affinity graph is divided into 8 feature sub-graphs by the spectral clustering algorithm, and each sub-graph contains a user group having similar behavior mode.

[0125] Feature sequence processing is performed in the feature sub-graphs. Taking a sub-graph containing 100,000 users as an example, the feature sequences are arranged according to the time order, and the sequences are divided into the query

feature sequence and the key-value pair feature sequence on the basis of the feature similarity threshold of 0.7. The time-order causal mask matrix with the size of 100,000 x 100,000 is constructed, only historical information before the current moment is reserved in the matrix, and the information of the future moment is shielded. The mask matrix is applied to an interaction result of the query sequence and the key-value pair sequence, thus obtaining the initial attention weight.

[0126] The spatial distance matrix of the sample features in the sub-graphs is computed. For each pair of samples, the Euclidean distance is computed based on their feature vectors, and the distance matrix with the size of 100,000 x 100,000 is constructed. The spatial distribution entropy is computed based on the distance matrix, and the entropy value reflects the uniformity degree of feature distribution. For example, if the spatial distribution entropy of a certain sub-graph is 0.75, it is indicated that the feature distribution is relatively uniform; and if the entropy value of the other sub-graph is 0.45, it is indicated that the feature distribution is relatively concentrated. The initial attention weight is adjusted according to the entropy value: for the sub-graphs which are uniformly distributed, the initial weight is kept; and for the sub-graphs which are intensively distributed, the attention weight of a local region is increased. The obtained adaptive attention weight is used for performing weighting on the key-value pair feature sequence.

[0127] In order to retain the location information of the features, the location encoding vector is constructed. The dimension of the location code is the same as that of the features and both are 256. The location code is added with the weighted feature sequences to obtain the local feature representation of each sub-graph. For example, the local feature representation of a certain sub-graph captures the common feature of the user group in recent shopping behaviors.

[0128] The four-layer multilayer perceptron network is constructed, and the numbers of neurons at each layer are 512, 256, 128, and 64. The splicing vector of the local feature representation of two adjacent feature sub-graphs is inputted, and forward transmission is performed through the network to obtain the association weight between the sub-graphs. For example, the association weight between the sub-graph representing a young user group and the sub-graph representing a middle-aged user group is 0.6, which indicates that shopping behaviors of the two groups have certain relevancy. Weighted summation is performed on the local feature representations of all the sub-graphs based on the association weight, thus obtaining the global feature

representation with the dimension of 256.

[0129] Self-calibration processing is performed on the global feature representation. Firstly, the mean value and the standard deviation on the feature dimension are computed. The dynamic normalization network which contains two  
 5 full-connection layers is constructed, and the numbers of neurons are 128 and 256 respectively. The global feature representation is inputted into the network to obtain the feature transformation coefficient and the feature offset. After performing standardization processing on the global feature representation, the global feature representation is multiplied by the transformation coefficient and the result is added to  
 10 the offset to obtain the transformation feature. The transformation feature is added to the original global feature representation to obtain the self-calibration feature.

[0130] Finally, layer normalization operation is performed on the self-calibration feature, and nonlinear transformation is performed through the ReLU activation function to obtain the final deep feature representation. The dimension of the deep  
 15 feature representation is kept to be 256, and each dimension contains the information obtained by performing multilayer transformation and fusion on the original feature. The features not only retain the behavior features of the local user group, but also contain the association information between different groups, thereby providing high-quality feature representations for subsequent feature selection and prediction  
 20 tasks.

[0131] In this embodiment, by the hierarchical attention mechanism and feature sub-graph division, the local relevancy between the samples can be better captured, and the discrimination capability for the feature representation is improved. Due to the causal attention in the sub-graphs, effective utilization of the time-order information is  
 25 ensured, and information leakage is avoided; and the spatial distribution entropy is added to perform adaptive adjustment on the attention weights, so that the attention distribution is more reasonable and the contributions of important features can be highlighted. Through the association weight learning between the feature sub-graphs, effective fusion of multi-scale features is realized; and the self-calibration module  
 30 enhances nonlinear expression capability of the features by means of dynamic normalization and residual connection, and keeps the original information. Through this design, the generalization performance of the model is improved, and the extracted features are more robust.

[0132] In an optional embodiment, the dividing the feature sequence of the

feature sub-graph into a query feature sequence and a key-value pair feature sequence, constructing a time-order causal mask matrix, and shielding an interaction result of the query feature sequence and the key-value pair feature sequence by the time-order causal mask matrix to obtain an initial attention weight only containing historical  
 5 feature information includes:

[0133] based on the similarity of adjacent time step features in the feature sub-graphs, grouping the feature sequence of the feature sub-graphs to obtain the query feature sequence and the key-value pair feature sequence, classifying a time step feature with the similarity of the adjacent time step features greater than a dynamic  
 10 threshold determined based on the global statistical characteristic of the feature sequence into the query feature sequence, classifying the rest of the features into the key-value pair feature sequence, so as to obtain the feature classification result of the query feature sequence and the key-value pair feature sequence;

[0134] based on the feature classification result, computing the time-order association intensity of the query feature sequence and the key-value pair feature sequence through the recurrent neural network, a hidden state of the recurrent neural network containing long short term dependence information of the query feature sequence and the key-value pair feature sequence, so as to obtain the time-order association intensity of the query feature sequence and the time-order association  
 20 intensity of the key-value pair feature sequence;

[0135] constructing a first mask matrix based on the time-order location relationship of the feature sequence to serve as a time-order causal mask matrix, transforming the time-order association intensity of the query feature sequence and the time-order association intensity of the key-value pair feature sequence into soft mask  
 25 weights to obtain a second mask matrix; and

[0136] performing matrix multiplication operation on the query feature sequence and the key-value pair feature sequence to obtain an original attention score; performing addition operation on the original attention score and the first mask matrix to obtain a first mask result, performing element-level multiplication operation on the  
 30 first mask result and the second mask matrix to obtain an attention score obtained after dual mask; and performing normalization processing on the attention score obtained after dual mask to obtain the initial attention weight only containing the historical feature information.

[0137] The soft mask weight specifically refers to a weight allocation mode based

on dynamic adjustment of importance of the features, and is used for emphasizing the contribution of different features to overall computation during feature sequence processing. By combining the time-order association intensity of the features and other context information, the weights between 0 and 1 are generated and used for representing a degree to which a certain feature is to be concerned in current computation.

[0138] The soft mask weight is continuously adjustable; and compared with a hard mask (completely shielded or completely passed), the soft mask weight can partially reduce or enhance the influence of specific features in a flexible mode. The dynamic weights can adaptively allocate attention resources in the operation process of the model according to the characteristics of data, so that the time-order relationship and dependency between the features can be captured more accurately.

[0139] In a specific embodiment, firstly, the feature sequences in the feature sub-graphs are dynamically divided. The similarity degree of the features is measured by computing the cosine similarity between adjacent time step features. Specifically, for any two adjacent time step features in the feature sequences, their cosine similarity values are computed. Moreover, a dynamic threshold is set based on the mean value and standard deviation statistical characteristic of the feature sequence. When the similarity of the adjacent time step features is greater than the dynamic threshold, these features are allocated into the query feature sequence, and the rest features are allocated into the key-value pair feature sequence. For example, for a feature sequence with the length of 10, if the feature similarity of the 2<sup>nd</sup>-4<sup>th</sup> time step and the 7<sup>th</sup>-8<sup>th</sup> time step is greater than a threshold of 0.8, the features of these time steps are allocated into the query feature sequence, and the features of the rest time steps are allocated into the key-value pair feature sequence.

[0140] Secondly, the long short term memory network is used for processing the query feature sequence and the key-value pair feature sequence respectively, and the time-order association information of the query feature sequence and the key-value pair feature sequence is extracted. The long short term memory network selectively retains and updates the feature information through the gating mechanism, and its hidden state contains the long-term dependency and the short-term relevancy of the feature sequence. For the query feature sequence, the hidden state output of the network represents its time-order association intensity; and for the key-value pair feature sequence, the hidden state output of the network represents its time-order

association intensity.

[0141] Thirdly, the dual mask mechanism is constructed. In a first mask, the causal mask matrix is constructed based on the time-order location relationship of the features, it is ensured that in the current time step, it can only visit historical information. In specific implementation, the value of the upper triangular region is set to be negative infinity, and the value of the lower triangular region is set to be 0. In a second mask, the time-order association intensity of the query sequence and the key-value pair sequence is transformed into the soft mask weight between 0 and 1 so as to form the soft mask matrix.

[0142] Fourthly, the attention weight is computed. The matrix multiplication is performed on the query feature sequence and the key-value pair feature sequence to obtain the original attention score, and the original attention score is added to the causal mask matrix to obtain a first mask result. The first mask result and the soft mask matrix are subjected to element-by-element multiplication to obtain the attention score after dual mask. The softmax normalization is performed on this score to obtain the final attention weight. This attention weight only contains historical time-order information and can be used for subsequent feature fusion.

[0143] By way of example, taking e-commerce user behavior sequence analysis as an example, the user behavior sequence in the feature sub-graph is processed. Firstly, the similarity of adjacent time step features in the feature sequence is computed. It is assumed to be the behavior sequence of a certain user in 30 consecutive days, the daily behavior feature vector dimension is 256. For the behavior features of any two adjacent days, their cosine similarity value is computed.

[0144] The dynamic threshold is determined based on the global statistical characteristic of the feature sequence. The mean value and the standard deviation of the similarity of all adjacent time step features in 30 days are counted, for example, the mean value is 0.65, and the standard deviation is 0.15. The mean value is added to 0.5 times of standard deviation to serve as the dynamic threshold, and the threshold in this example is 0.725. The similarity value of the adjacent time steps is compared with the threshold, the time step features with the similarity greater than the threshold are allocated into the query feature sequence, and the other features are allocated into the key-value pair feature sequence.

[0145] For example, if the behavior feature similarity of the 3<sup>rd</sup> day and the 4<sup>th</sup> day is 0.78, which is greater than a threshold of 0.725, the features of the 4<sup>th</sup> day are

allocated into the query feature sequence; and if the feature similarity of the 15<sup>th</sup> day and the 16<sup>th</sup> day is 0.68, which is less than the threshold, the features of the 16<sup>th</sup> day are allocated into the key-value pair feature sequence. After allocation, the query feature sequence containing 12 time steps and the key-value pair feature sequence containing  
 5 18 time steps can be obtained.

[0146] The bidirectional long short term memory network (Bi-LSTM) is constructed as the recurrent neural network for computing the time-order association intensity. The network contains a dual-layer Bi-LSTM structures, and the number of hidden layer units is 128. The query feature sequence is inputted into this network, and  
 10 the hidden state of the network contains the long short term dependence information of the feature sequence. For the query feature sequence of 12 time steps, a time-order association intensity matrix with the dimension of 12 can be obtained. Similarly, the key-value pair feature sequence of 18 time steps is inputted into the network to obtain a time-order association intensity matrix with the dimension of 18 x 128.

[0147] The time-order causal mask matrix is constructed based on the time-order location relationship of the feature sequence. The matrix size is 30 x 30, it corresponds to a complete 30-day sequence. In the matrix, only the information of the current time step and the previous time step is reserved, and the location of the later time step is marked to be negative infinity and treated as a first mask matrix. For example, for the  
 20 prediction of the 10<sup>th</sup> day, only the feature information of the 1<sup>st</sup> day to the 10<sup>th</sup> day is taken into account, and the information of the 11<sup>th</sup> day to the 30<sup>th</sup> day is shielded by the mask.

[0148] The time-order association intensity of the query feature sequence and the key-value pair feature sequence is transformed into the soft mask weights. Softmax  
 25 normalization processing is performed on the association intensity to obtain the second mask matrix. This matrix reflects the association degree between different time step features, and the larger the association intensity is, the larger the mask weight corresponding to the time step is.

[0149] Matrix multiplication operation is performed on the query feature  
 30 sequence and the key-value pair feature sequence to obtain an original attention score matrix. The matrix dimension is the query sequence length multiplied by the key-value pair sequence length, that is, 12 x 18. This matrix is added to the first mask matrix to realize time-order causal mask, and it is ensured that only historical information is used during prediction. Then, the mask result and the second mask matrix are subjected to

element-level multiplication operation to obtain a dual-mask attention score based on both the time-order causal relationship and the feature association intensity.

[0150] Finally, Softmax normalization processing is performed on the attention score after dual mask to obtain the initial attention weights. These weights reflect the association degree of each time step and key value in the query sequence to the historical time step in the sequence, for example, the attention weight of the query feature of the 10<sup>th</sup> day and the key value of the 8<sup>th</sup> day to the feature is 0.15, which indicates that the features of the two time steps have relatively high relevancy.

[0151] In this embodiment, the query sequence and the key-value pair sequence are adaptively divided by the dynamic feature sequence division mechanism according to the feature similarity, thereby improving the discrimination of the feature representation and enhancing the sensing capability of the model to important features; the dual-mask mechanism is adopted, thereby guaranteeing strict time-order causality, and the flexible feature selection mechanism is added through the soft masking weight, which effectively inhibits the interference of irrelevant features and improves the robustness of the model; and the time-order association intensity is extracted based on the recurrent neural network, the long short term dependence information of the feature sequence is fully utilized, thus the modeling capability of the model to the time-order mode is enhanced, and the time-order consistency of the feature representation is improved.

[0152] In an optional embodiment, the constructing an information bottleneck feature filter, and filtering the deep feature representation by maximizing task related information and minimizing redundant information to obtain a key feature subset includes:

[0153] constructing the information bottleneck encoder, the information bottleneck encoder including a three-layer perceptron structure, a normalization layer being connected behind each layer of perceptron structure, inputting the deep feature representation into the information bottleneck encoder, and generating probability distribution parameters of a compression feature space at an output layer of the information bottleneck encoder; performing code mapping on the deep feature representation based on the probability distribution parameters to obtain a compression feature representation;

[0154] constructing a mutual information assessment module, the mutual information assessment module containing the three-layer perceptron structures,

inputting a splicing vector of the compression feature representation and a tag of the target task into the mutual information assessment module, and outputting mutual information amount between the compression feature representation and the target task;

5        [0155] computing an autocorrelation matrix of the compression feature representation, each element in the autocorrelation matrix being obtained by dividing a product that is obtained by subtracting a corresponding mean value from a corresponding feature vector by a corresponding standard deviation; computing feature redundancy based on a difference value between the autocorrelation matrix and a unit  
10    matrix;

      [0156] constructing a feature selection network, the feature selection network including a first fully-connected layer and a second fully-connected layer, inputting the compression feature representation into the first fully-connected layer, inputting the output of the first fully-connected layer into the second fully-connected layer after  
15    processing through the activation function, processing the output of the second fully-connected layer through a gating function to obtain a gating value, and performing element-by-element multiplication on the gating value and the compression feature representation to obtain an initial filter feature;

      [0157] adjusting the output of the gating function based on temperature  
20    parameters, the temperature parameters decreasing progressively along with the number of training rounds based on a preset decreasing amount, and performing element-by-element multiplication on the adjusted gating value and the compression feature representation to obtain an adjusted filter feature;

      [0158] performing linear combination of an opposite number of the mutual  
25    information amount, the feature redundancy and a norm of the adjusted filter feature to construct a target function;

      [0159] optimizing the target function by iteration to obtain an optimization parameter of the information bottleneck encoder and an optimization parameter of the feature selection network, so as to obtain an optimized information bottleneck encoder  
30    and an optimized feature selection network; and

      [0160] inputting the deep feature representation into the optimized information bottleneck encoder to obtain the optimized compression feature representation, and inputting the optimized compression feature representation into the optimized feature selection network to obtain the key feature subset.

[0161] In a specific embodiment, in the construction process of the information bottleneck feature filter, it is needed to construct the information bottleneck encoder. The encoder utilizes the three-layer perceptron structure, the numbers of neurons in each layer of perceptron are 1,024, 512 and 256 respectively, and ReLU is used as the  
 5 activation function. The normalization layer is connected behind each layer of perceptron for accelerating the training convergence. The inputted deep feature representation dimension is 2048, and two groups of parameters, namely, the mean vector and the variance vector, are outputted after being processed by the encoder, which are respectively used for constructing probability distribution of the compressed  
 10 feature space. Based on a reparameterization technology, sampling is performed from the probability distribution to obtain the compressed feature representation with the dimension of 256.

[0162] The mutual information assessment module also adopts the three-layer perceptron structure, and the numbers of neurons are 256, 512 and 1. The compressed  
 15 feature representation and the task tag are spliced as input, and the tag is in a one-hot encoding form. By means of nonlinear transformation, a scalar value is finally outputted to represent mutual information between the compressed feature and the task.

[0163] In the computation of feature redundancy, firstly, it is needed to perform  
 20 standardized processing on the compression feature. The mean value and standard deviation of each feature dimension are computed, the mean value is subtracted from the features and the result is divided by the standard deviation. The inner product between the standardized features is computed to obtain the autocorrelation matrix. The quadratic sum of the difference value between the matrix and the unit matrix is the  
 25 feature redundancy tolerance.

[0164] The feature selection network contains two fully-connected layers, the numbers of neurons in the first layer is 512, and the ReLU activation function is used; the numbers of neurons in the second layer are the same as the compression feature dimension, namely 256, and a Sigmoid function is used as the gating function. The  
 30 network inputs the compression feature representation, and outputs the gating value between 0 and 1. The gating value and the compression features are subjected to element-by-element multiplication to realize soft selection of the features.

[0165] In order to make the feature selection sparser, the temperature parameter is added to adjust the gating function. The initial value of the temperature parameter is set

to be 1.0, and the temperature parameter is decreased by 0.1 every 100 rounds of training until the minimum value is 0.1. The smaller the temperature parameter is, the closer the gating value is to 0 or 1, and the sparser the feature selection is.

[0166] The target function is composed of three parts: mutual information loss, redundancy loss and sparsity loss. The mutual information loss is an inverse number of an estimated mutual information value; the redundancy loss is the feature redundancy; and the sparsity loss is an L1 norm of the filter features after adjusting. The weight ratio of the three losses is 1:0.5:0.1.

[0167] In the training process, an Adam optimizer is utilized to perform parameter updating, and the learning rate is set to be 0.001. Each minibatch contains 128 samples, and it is totally trained in 50 epochs. After training, the optimized encoder and the feature selection network are utilized to process new deep features to obtain the final key feature subset.

[0168] In this embodiment, the feature compression and information extraction are implemented by the information bottleneck encoder, so that the feature dimension is effectively reduced, the compactness and effectiveness of the feature representation are improved, and moreover, the key information related to the task is reserved; based on dual constraints of the mutual information assessment and redundancy tolerance, the features obtained by filtering are ensured to be highly related to the task and have relatively low information redundancy, so that the efficiency and distinguishability of the feature representation are improved; and adaptive soft selection of the features is realized by the learnable feature selection network in combination with the temperature adjustment mechanism, so that the microminiaturization of the selection process is kept, sparse feature representation can be obtained, and the interpretability and generalization capability of the model are improved.

[0169] In an optional embodiment, the according to the feature weight vector and the prediction result, computing the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result includes:

[0170] based on the feature weight vector and the prediction result, computing the association coefficient between each pair of features in the feature weight vector, so as to generate the feature association matrix; constructing a feature dependency graph based on the feature association matrix, treating each feature in the key feature subset as a node in the feature dependency graph, and treating the association coefficient in

the feature association matrix serves as a weight of an edge between the corresponding nodes;

[0171] constructing a graph Laplacian matrix according to the feature dependency graph, treating the sum of the edge weight of each node and the edge weight of the  
5 corresponding connected node as a diagonal element value of the graph Laplacian matrix, and treating a negative value of the edge weights between the nodes as a non-diagonal element value of the graph Laplacian matrix;

[0172] computing a Jacobian matrix of the prediction result about the key feature subset to obtain an initial gradient; treating the graph Laplacian matrix as a structured  
10 regularization item, and performing weighted combination on the structured regularization item and the initial gradient to obtain the structured gradient under the feature dependence relationship;

[0173] computing a structured gradient amplitude of each feature in the key feature subset, an interaction gradient value of each feature and other features, and  
15 centrality measurement of each feature in the feature dependency graph; and

[0174] performing weighted combination on the structured gradient amplitudes, the interaction gradient values and the centrality measurements to obtain the feature contribution; performing normalization processing on the feature contribution to obtain a standardized contribution of each feature in the key feature subset to the prediction  
20 result; and performing importance sorting on the features in the key feature subset based on the standardized contribution to generate the analysis result.

[0175] In a specific embodiment, firstly, the feature weight vector and the prediction result data are acquired. Taking a data set containing 10 key features as an example, the features include environmental parameters such as temperature, humidity,  
25 and pressure. The prediction result is a quality index of a certain product.

[0176] Relevancy analysis is performed on feature pairs in the feature weight vector. The feature association matrix is obtained by computing the Pearson association coefficients. For example, the association coefficient of the temperature and the humidity is 0.7, which represents a relatively strong positive association  
30 relationship. After the association coefficients of all the feature pairs are computed, the complete feature association matrix can be obtained.

[0177] The feature dependency graph is constructed based on the feature association matrix. Each feature is treated as the node in the graph, and the association coefficient is taken as the weight of the edges connected to the node. For example, the

weight of the edge between a temperature node and a humidity node is 0.7. For the feature pairs with the association coefficients less than the threshold, the edges in the graph cannot be connected, in order to highlight important feature dependence relationship.

5        [0178] The graph Laplacian matrix is constructed, the sum of the edge weight of each node and the edge weight of the connected node is computed as the diagonal element. For example, the sum of the edge weights of the temperature node and other nodes is 2.5, and it is treated as the diagonal element. A negative value of the weight of the edge between the corresponding nodes is taken as the non-diagonal element, and  
10       the corresponding locations of the nodes without edge connections are filled with zero.

      [0179] The initial gradient of the prediction result for each feature is computed. Tiny disturbance is performed on each feature by the finite difference method, and the change of the prediction result is observed. Weighted combination is performed on the graph Laplacian matrix and the initial gradients to obtain the structured gradients under  
15       the feature dependence relationship. The weighting coefficients can be adjusted according to actual demands and used for balancing the influence of feature independence and dependency.

      [0180] The structured gradient amplitude of each feature is computed to represent the direct influence degree of the feature on the prediction result. The interaction  
20       gradient between the features is computed to represent the intensity of the feature synergistic effect. The centrality of the feature in the dependency graph is computed, including indexes such as degree centrality and proximity centrality, to represent the importance degree of the feature in the whole feature network.

      [0181] Weighted combination is performed on the above three indexes to obtain  
25       the feature contribution. For example, the weight of the structured gradient amplitude can be set to be 0.5, the weight of the interaction gradient is set to be 0.3, and the weight of the centrality is set to be 0.2. Normalization processing is performed on the feature contribution to make the sum of the contributions of all the features be 1. Finally, the features are sorted according to the standardized contribution to generate  
30       the analysis result report.

      [0182] In this embodiment, the feature dependency graph and the graph Laplacian matrix are constructed to effectively capture the complex dependence relationship between the features, thus the problem that the feature interaction effect is ignored in a traditional feature importance analysis method can be avoided, and the accuracy of

feature contribution assessment is improved; by the structured gradient method, the direct influence, interaction effect and network centrality of the features are taken into full account, thus multi-dimensional feature importance assessment is realized, the analysis result is more comprehensive and reliable, and feature selection and optimization in practical application can be better guided; and by standardized processing and sorting, a visual and clear feature importance sorting result is obtained, which facilitates understanding and application, and ensuring the comparability of analysis results in different scenarios, and therefore, a reliable basis is provided for subsequent model optimization and decision making.

10 [0183] FIG. 2 is a schematic structural diagram of a Transformer-based modeling and analysis system for structured data according to an embodiment of the present invention, as shown in FIG. 2, the system includes:

[0184] a first unit, configured to receive structured original data, divide the structured original data into a numerical feature and a categorical feature, execute  
15 piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and execute adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; perform adaptive weight allocation by a learnable gating mechanism, and fuse the normalized main feature with the continuous value mapping feature to obtain a fusion feature; and  
20 perform dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set;

[0185] a second unit, configured to input the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity  
25 matrix of sample features, constructing a feature affinity graph on the basis of the similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extract local feature representation by a causal attention mechanism, and integrate a plurality of local feature representation through a cross-sub-graph attention fusion mechanism to  
30 obtain a global feature representation; and receive the global feature representation by the self-calibration module, and perform feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation; and

[0186] a third unit, configured to construct an information bottleneck feature filter, and filter the deep feature representation by maximizing task related information and

minimizing redundant information to obtain a key feature subset; input the key feature subset into a dual-branch network containing a main prediction branch and an auxiliary reconstruction branch, the main prediction branch processing the key feature subset based on the attention mechanism to obtain a feature weight vector, and generating a prediction result according to the feature weight vector, and the auxiliary reconstruction branch providing a regularization constraint by reconstructing the key feature subset; and according to the feature weight vector and the prediction result, compute the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result.

10 [0187] In a third aspect,

[0188] an electronic device provided by the embodiment of the present invention includes:

[0189] a processor; and

[0190] a memory configured to store processor executable instructions;

15 [0191] the processor is configured to call instructions stored in the memory to execute the abovementioned method.

[0192] In a fourth aspect,

[0193] 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.

20

[0194] 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.

25

[0195] 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.

30

## CLAIMS

We claim:

1. A Transformer-based modeling and analysis method for structured data, comprising:  
 receiving structured original data, dividing the structured original data into a  
 5 numerical feature and a categorical feature, executing piecewise linear transformation on  
 the numerical feature by a dual-stream feature encoder to obtain a normalized main feature,  
 and executing adaptive quantum encoding on the categorical feature to obtain a continuous  
 value mapping feature; performing adaptive weight allocation by a learnable gating  
 mechanism, and fusing the normalized main feature with the continuous value mapping  
 10 feature to obtain a fusion feature; performing dynamic sampling on the basis of an  
 importance score of the fusion feature to generate an enhanced training sample set;  
 inputting the enhanced training sample set into a feature extraction network which  
 contains a hierarchical attention module and a self-calibration module, the hierarchical  
 attention module computing a similarity matrix of sample features, constructing a feature  
 15 affinity graph on the basis of the similarity matrix, and dividing the feature affinity graph  
 into a plurality of feature sub-graphs through a graph segmentation algorithm; in each  
 feature sub-graph, extracting a local feature representation by a causal attention mechanism,  
 and integrating a plurality of local feature representations through a cross-sub-graph  
 attention fusion mechanism to obtain a global feature representation; receiving the global  
 20 feature representation by the self-calibration module, and performing feature nonlinear  
 transformation by residual connection and dynamic normalization to obtain a deep feature  
 representation;  
 constructing an information bottleneck feature filter, and filtering the deep feature  
 representation by maximizing task related information and minimizing redundant  
 25 information to obtain a key feature subset; inputting the key feature subset into a  
 dual-branch network containing a main prediction branch and an auxiliary reconstruction  
 branch, the main prediction branch processing the key feature subset based on the attention  
 mechanism to obtain a feature weight vector, and generating a prediction result according  
 to the feature weight vector, and the auxiliary reconstruction branch providing a  
 30 regularization constraint by reconstructing the key feature subset; and according to the  
 feature weight vector and the prediction result, computing the contribution of each feature  
 in the key feature subset to the prediction result by a structured gradient method to generate  
 an analysis result of a user purchase intention prediction.
2. The method according to claim 1, wherein the receiving structured original data,

dividing the structured original data into a numerical feature and a categorical feature, executing piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and executing adaptive quantum encoding on the categorical feature to obtain a continuous value mapping feature; performing adaptive  
 5 weight allocation by a learnable gating mechanism, and fusing the normalized main feature with the continuous value mapping feature to obtain a fusion feature; performing dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set comprises:

receiving the structured original data, and dividing the structured original data into the  
 10 numerical feature and the categorical feature; inputting the numerical feature and the categorical feature into the dual-stream feature encoder for feature processing, the dual-stream feature encoder comprising a feature main stream and a feature mapping stream;

in the feature main stream, counting the data distribution of the numerical feature to  
 15 obtain data density, determining a location of an optimal segmentation point based on the data density, dividing the numerical feature into a plurality of segmentation intervals according to the optimal segmentation point, constructing a linear transformation parameter matrix in each segmentation interval, computing a feature difference value between adjacent segmentation intervals, determining an optimized linear transformation parameter  
 20 matrix of each segmentation interval by minimizing the feature difference value, and executing linear transformation on the numerical feature of each segmentation interval by the optimized linear transformation parameter matrix, so as to obtain the normalized main feature;

in the feature mapping stream, mapping each category value in the categorical feature  
 25 into a corresponding quantum ground state, allocating an initial amplitude parameter and a phase parameter to each quantum ground state, constructing a quantum gate sequence containing a rotating gate and a control gate, sequentially acting the quantum gate sequence on the quantum ground states, performing quantum state transformation to obtain a transformed quantum state, computing an entropy value of the transformed quantum state,  
 30 and projecting the transformed quantum state to a continuous space based on the entropy value to obtain the continuous value mapping feature;

computing an association coefficient between the normalized main feature and the continuous value mapping feature in each dimension respectively, combining the association coefficients to construct the feature association coefficient matrix, training a

preset attention network based on the feature association coefficient matrix to obtain an importance vector of the feature dimension, and performing weighted combination on the normalized main feature and the continuous value mapping feature through the importance vector to obtain the fusion feature; and

5        computing the contribution of each dimension of the fusion feature to a target task, generating the feature importance score of each dimension based on the contribution, constructing a sampling probability distribution function according to the feature importance score, and performing importance sampling on training samples by the sampling probability distribution function to generate the enhanced training sample set.

10        3. The method according to claim 2, wherein the counting the data distribution of the numerical feature to obtain data density, determining a location of an optimal segmentation point based on the data density comprises:

      computing the width of a basic interval based on a interquartile range of the numerical feature and the number of samples, the numerical feature comprising user age; dividing the value range of the numerical feature into a plurality of basic statistical intervals based on the width of the basic interval, and counting the number of samples in each basic statistical interval to obtain an interval frequency sequence;

      counting a frequency ratio of adjacent basic statistical intervals in the interval frequency sequence to obtain the interval similarity, and merging the adjacent basic statistical intervals with the interval similarity greater than a preset merging threshold into a new statistical interval to obtain a merged frequency distribution sequence;

      computing accumulated frequency distribution according to the merged frequency distribution sequence, and generating a data density function based on the accumulated frequency distribution; selecting candidate segmentation locations from the merged frequency distribution sequence, computing the number of samples and the numerical mean value of the intervals on two sides of each candidate segmentation location, and computing an interval variance value of the candidate segmentation locations by means of the number of samples and the numerical mean value;

      selecting a location with the maximum interval variance value in the current intervals to be segmented to serve as a segmentation point to be determined, computing a ratio of the interval variance value of the segmentation point to be determined to the overall variance of the numerical feature, and determining a variance contribution ratio of the segmentation point to be determined; and

      computing local distribution skewness of the intervals on the two sides of the

segmentation point to be determined, when the difference value of the local distribution skewness is greater than a preset skewness threshold and the variance contribution ratio is greater than a preset segmentation threshold, determining the segmentation point to be determined to be the optimal segmentation point, and repeating the steps to determine the  
 5 locations of all the optimal segmentation points.

4. The method according to claim 1, wherein the inputting the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity matrix of sample features, constructing a feature affinity graph on the basis of the similarity  
 10 matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extracting a local feature representation by a causal attention mechanism, and integrating a plurality of local feature representations through a cross-sub-graph attention fusion mechanism to obtain a global feature representation; receiving the global feature representation by the self-calibration  
 15 module, and performing feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation comprises:

inputting the enhanced training sample set into a convolutional neural network to obtain an initial feature map, computing a dot product of any two sample features in the initial feature map, and dividing the dot product by a module length product of a  
 20 corresponding feature vector to obtain the feature similarity matrix; constructing the feature affinity graph based on the feature similarity matrix, a vertex of the feature affinity graph corresponding to the sample feature, an edge weight of the feature affinity graph corresponding to a similarity value in the feature similarity matrix, and dividing the feature affinity graph into a plurality of feature sub-graphs by a spectral clustering algorithm;

25 in each feature sub-graph, dividing the feature sequence of the feature sub-graph into a query feature sequence and a key-value pair feature sequence, constructing a time-order causal mask matrix, and shielding an interaction result of the query feature sequence and the key-value pair feature sequence by the time-order causal mask matrix to obtain an initial attention weight only containing historical feature information; computing a spatial  
 30 distance matrix of the sample features in each feature sub-graph, computing a spatial distribution entropy of the feature sub-graph based on the spatial distance matrix, and performing adaptive adjustment on the initial attention weight according to the spatial distribution entropy to obtain the adaptive attention weight; weighting the key-value pair feature sequence based on the adaptive attention weight, and superposing the location

encoding information to obtain the local feature representation of each feature sub-graph;

based on a pre-constructed multilayer perceptron network, inputting a splicing vector of the local feature representation of adjacent feature sub-graphs, performing forward transmission through the multilayer perceptron network to obtain an association weight  
 5 between the feature sub-graphs, and performing weighted summation on the local feature representation based on the association weight to obtain the global feature representation;

computing a mean value and a standard deviation of the global feature representation in the feature dimension; based on a pre-constructed dynamic normalization network, inputting the global feature representation into the dynamic normalization network, and  
 10 performing forward transmission to obtain a feature transformation coefficient and a feature offset; subtracting the mean value from the global feature representation and dividing the result by the standard deviation to obtain a standardized feature, multiplying the standardized feature by the feature transformation coefficient, and adding the feature offset to obtain the transformation feature, adding the transformation feature and the global  
 15 feature representation to obtain a self-calibration feature; and

performing layer normalization processing on the self-calibration feature, and performing nonlinear transformation through a ReLU activation function to obtain the deep feature representation.

5. The method according to claim 4, wherein the dividing the feature sequence of the  
 20 feature sub-graph into a query feature sequence and a key-value pair feature sequence, constructing a time-order causal mask matrix, and shielding an interaction result of the query feature sequence and the key-value pair feature sequence by the time-order causal mask matrix to obtain an initial attention weight only containing historical feature information comprises:

25 the feature sub-graph comprising e-commerce user behaviors of a user over the past 30 consecutive days, forming a feature sequence of the user behaviors, computing similarity of adjacent time step features in the feature sequence, grouping the feature sequence to obtain the query feature sequence and the key-value pair feature sequence, classifying a time step feature with the similarity of the adjacent time step features greater  
 30 than a dynamic threshold determined based on the global statistical characteristic of the feature sequence into the query feature sequence, classifying the rest of the features into the key-value pair feature sequence, so as to obtain the feature classification result of the query feature sequence and the key-value pair feature sequence;

based on the feature classification result, computing the time-order association

intensity of the query feature sequence and the key-value pair feature sequence through the recurrent neural network, a hidden state of the recurrent neural network containing long short term dependence information of the query feature sequence and the key-value pair feature sequence, so as to obtain the time-order association intensity of the query feature sequence and the time-order association intensity of the key-value pair feature sequence;

constructing a first mask matrix based on the time-order location relationship of the feature sequence to serve as a time-order causal mask matrix, transforming the time-order association intensity of the query feature sequence and the time-order association intensity of the key-value pair feature sequence into soft mask weights to obtain a second mask matrix; and

performing matrix multiplication operation on the query feature sequence and the key-value pair feature sequence to obtain an original attention score; performing addition operation on the original attention score and the first mask matrix to obtain a first mask result, performing element-level multiplication operation on the first mask result and the second mask matrix to obtain an attention score obtained after dual mask; and performing normalization processing on the attention score obtained after dual mask to obtain the initial attention weight only containing the historical feature information.

6. The method according to claim 1, wherein the constructing an information bottleneck feature filter, and filtering the deep feature representation by maximizing task related information and minimizing redundant information to obtain a key feature subset comprises:

constructing the information bottleneck encoder, the information bottleneck encoder comprising a three-layer perceptron structure, a normalization layer being connected behind each layer of perceptron structure, inputting the deep feature representation into the information bottleneck encoder, and generating probability distribution parameters of a compression feature space at an output layer of the information bottleneck encoder; performing code mapping on the deep feature representation based on the probability distribution parameters to obtain a compression feature representation;

constructing a mutual information assessment module, the mutual information assessment module containing the three-layer perceptron structures, inputting a splicing vector of the compression feature representation and a tag of the target task into the mutual information assessment module, and outputting mutual information amount between the compression feature representation and the target task;

computing an autocorrelation matrix of the compression feature representation, each

element in the autocorrelation matrix being obtained by dividing a product that is obtained by subtracting a corresponding mean value from a corresponding feature vector by a corresponding standard deviation; computing feature redundancy based on a difference value between the autocorrelation matrix and a unit matrix;

5        constructing a feature selection network, the feature selection network comprising a first fully-connected layer and a second fully-connected layer, inputting the compression feature representation into the first fully-connected layer, inputting the output of the first fully-connected layer into the second fully-connected layer after processing through the activation function, processing the output of the second fully-connected layer through a  
10        gating function to obtain a gating value, and performing element-by-element multiplication on the gating value and the compression feature representation to obtain an initial filter feature;

         adjusting the output of the gating function based on temperature parameters, the temperature parameters decreasing progressively along with the number of training rounds  
15        based on a preset decreasing amount, and performing element-by-element multiplication on the adjusted gating value and the compression feature representation to obtain an adjusted filter feature;

         performing linear combination of an opposite number of the mutual information amount, the feature redundancy and a norm of the adjusted filter feature to construct a  
20        target function;

         optimizing the target function by iteration to obtain an optimization parameter of the information bottleneck encoder and an optimization parameter of the feature selection network, so as to obtain an optimized information bottleneck encoder and an optimized feature selection network; and

25        inputting the deep feature representation into the optimized information bottleneck encoder to obtain the optimized compression feature representation, and inputting the optimized compression feature representation into the optimized feature selection network to obtain the key feature subset.

7. The method according to claim 1, wherein the according to the feature weight  
30        vector and the prediction result, computing the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result comprises:

         based on the feature weight vector and the prediction result, computing the association coefficient between each pair of features in the feature weight vector, so as to generate the

feature association matrix; constructing a feature dependency graph based on the feature association matrix, treating each feature in the key feature subset as a node in the feature dependency graph, and treating the association coefficient in the feature association matrix serves as a weight of an edge between the corresponding nodes;

5        constructing a graph Laplacian matrix according to the feature dependency graph, treating the sum of the edge weight of each node and the edge weight of the corresponding connected node as a diagonal element value of the graph Laplacian matrix, and treating a negative value of the edge weights between the nodes as a non-diagonal element value of the graph Laplacian matrix;

10        computing a Jacobian matrix of the prediction result about the key feature subset to obtain an initial gradient; treating the graph Laplacian matrix as a structured regularization item, and performing weighted combination on the structured regularization item and the initial gradient to obtain the structured gradient under the feature dependence relationship;

      computing a structured gradient amplitude of each feature in the key feature subset, an  
15        interaction gradient value of each feature and other features, and centrality measurement of each feature in the feature dependency graph; and

      performing weighted combination on the structured gradient amplitudes, the interaction gradient values and the centrality measurements to obtain the feature contribution; performing normalization processing on the feature contribution to obtain a  
20        standardized contribution of each feature in the key feature subset to the prediction result; and performing importance sorting on the features in the key feature subset based on the standardized contribution to generate the analysis result of the user purchase intention prediction.

      8. A Transformer-based modeling and analysis system for structured data, applied to  
25        the method according to any one of claims 1 to 7, comprising:

      a first unit, configured to receive structured original data, divide the structured original data into a numerical feature and a categorical feature, execute piecewise linear transformation on the numerical feature by a dual-stream feature encoder to obtain a normalized main feature, and execute adaptive quantum encoding on the categorical feature  
30        to obtain a continuous value graphing feature; perform adaptive weight allocation by a learnable gating mechanism, and fuse the normalized main feature with the continuous value graphing feature to obtain a fusion feature; and perform dynamic sampling on the basis of an importance score of the fusion feature to generate an enhanced training sample set;

a second unit, configured to input the enhanced training sample set into a feature extraction network which contains a hierarchical attention module and a self-calibration module, the hierarchical attention module computing a similarity matrix of sample features, constructing a feature affinity graph on the basis of the similarity matrix, and dividing the  
 5 feature affinity graph into a plurality of feature sub-graphs through a graph segmentation algorithm; in each feature sub-graph, extract local feature representation by a causal attention mechanism, and integrate a plurality of local feature representation through a cross-sub-graph attention fusion mechanism to obtain a global feature representation; and receive the global feature representation by the self-calibration module, and perform  
 10 feature nonlinear transformation by residual connection and dynamic normalization to obtain a deep feature representation; and

a third unit, configured to construct an information bottleneck feature filter, and filter the deep feature representation by maximizing task related information and minimizing redundant information to obtain a key feature subset; input the key feature subset into a  
 15 dual-branch network containing a main prediction branch and an auxiliary reconstruction branch, the main prediction branch processing the key feature subset based on the attention mechanism to obtain a feature weight vector, and generating a prediction result according to the feature weight vector, and the auxiliary reconstruction branch providing a regularization constraint by reconstructing the key feature subset; and according to the  
 20 feature weight vector and the prediction result, compute the contribution of each feature in the key feature subset to the prediction result by a structured gradient method to generate an analysis result.

9. An electronic device, comprising:

a processor; and

25 a memory configured to store processor executable instructions; and

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