ARTIFICIAL INTELLIGENCE BIG DATA REAL-TIME PROCESSING AND ANALYSIS METHOD

TECHNICAL FIELD

The present invention relates to the technical field of big data processing, and in particular to an artificial intelligence (AI) big data real-time processing and analysis method.

BACKGROUND

With the advent of the big data era, a large amount of multi-source heterogeneous data has been generated in various fields such as finance, government affairs, and industry, and these data contain enormous value. As the core means of mining data value, big data real-time processing and analysis technology directly affects the timeliness and accuracy of decisions. However, existing big data processing and analysis methods have many defects: first, low real-time analysis efficiency. The traditional centralized collection architecture is difficult to cope with the concurrent collection of multi-source data, resulting in high data transmission and processing delays, which cannot meet the needs of real-time decision-making; second, insufficient analysis accuracy. A single analysis model is difficult to adapt to the complex features of multi-source heterogeneous data, and the mining of deep correlations and potential laws of data is not in-depth enough; third, limited analysis scope. Most analyses are aimed at structured data, with insufficient coverage of unstructured data (text, images, audio) and temporal data; fourth, weak data security and privacy protection capabilities. Data leakage or tampering is prone to occur during data transmission and processing, and there is a lack of effective protection mechanisms for user-sensitive information; fifth, low intelligence in data classification. Most adopt traditional rule-based classification methods, which are difficult to adapt to the dynamic changes of data features, resulting in low classification accuracy and poor flexibility.

To solve the above problems, some improved schemes have appeared in related technologies, such as adopting a distributed architecture to improve data processing efficiency and a single encryption algorithm to ensure data security, but there are still the following deficiencies: the insufficient coordination of the distributed architecture leads to limited improvement in the real-time performance of data collection and processing; a single analysis model cannot meet the analysis needs of multiple types of data, and the analysis accuracy and scope are still limited; the combination of encryption algorithms and privacy protection technologies is not close enough, making it difficult to achieve full-process security protection

for data processing; data classification still relies on manually set rules, with low intelligence and inability to achieve dynamic and accurate classification. Therefore, the present invention proposes an AI big data real-time processing and analysis method to solve the above-mentioned problems.

SUMMARY

Aiming at the deficiencies of the prior art, the present invention provides an AI big data real-time processing and analysis method, which solves the problems mentioned in the above background technology.

To achieve the above objectives, the present invention is realized through the following technical solutions: an AI big data real-time processing and analysis method includes the following steps:

S1: conducting real-time data collection and preprocessing, adopting a distributed collection architecture to collect multi-source heterogeneous big data, performing real-time cleaning and format standardization on the collected data, and generating a standardized real-time data stream;

S2: implementing intelligent encryption and secure transmission, adopting a "symmetric encryption + asymmetric encryption" hybrid encryption algorithm to encrypt the standardized real-time data stream, transmitting it to analysis nodes through a secure transmission protocol, and simultaneously performing desensitization on user-sensitive information in the data;

S3: carrying out multi-dimensional intelligent classification, extracting multi-dimensional features and performing intelligent classification on the encrypted and desensitized real-time data based on an improved deep learning classification model to obtain multi-category data subsets;

S4: performing AI real-time analysis, matching corresponding AI analysis models for different category data subsets to construct a multi-model fusion analysis engine, realizing real-time in-depth analysis of data and outputting preliminary analysis results; and

S5: conducting result feedback and optimization, collecting application feedback data of analysis results and performance data during the data processing and analysis process, dynamically adjusting preprocessing rules, classification model parameters, and AI analysis model weights, and realizing continuous optimization of processing and analysis strategies.

Preferably, in step S1, the distributed collection architecture comprises edge collection nodes, aggregation nodes, and a central processing node; the edge collection nodes are deployed at various data sources, adopting the stream processing framework Flink to realize real-time

collection and preliminary filtering of data; the aggregation node aggregates and converts the format of data collected by multiple edge nodes; the central processing node adopts a data cleaning algorithm to remove duplicate data, abnormal data, and missing value data, and completes data format standardization through schema mapping to generate a standardized real-time data stream.

Preferably, in step S2, the specific implementation process of the hybrid encryption algorithm is: adopting the AES-256 symmetric encryption algorithm to batch encrypt the standardized real-time data stream to generate encrypted data; adopting the RSA-2048 asymmetric encryption algorithm to encrypt the key of AES-256 to generate a key ciphertext; transmitting the encrypted data and the key ciphertext in a bound manner; the desensitization adopts a combination of dynamic mask desensitization and differential privacy, performing mask processing on user-sensitive information such as ID numbers, mobile phone numbers, and bank card numbers, and adding noise conforming to Gaussian distribution to user behavior data.

Preferably, in step S3, the improved deep learning classification model is a CNN-LSTM fusion model based on an attention mechanism, which comprises a feature extraction layer, an attention enhancement layer, and a classification output layer; the feature extraction layer extracts local features of data through CNN and temporal features of data through LSTM; the attention enhancement layer assigns weights to the extracted local features and temporal features to strengthen the influence of key features; the classification output layer adopts a softmax activation function to realize multi-dimensional classification of data, and the classification dimensions include data source dimension, data type dimension, data importance dimension, and user demand dimension.

Preferably, in step S4, the multi-model fusion analysis engine comprises a regression analysis model, a clustering analysis model, an association rule mining model, and a deep learning prediction model; for structured data subsets, the regression analysis model and the association rule mining model are adopted for correlation analysis and trend prediction; for unstructured data subsets (text, images, audio), the deep learning prediction model is adopted for content analysis and semantic understanding; for temporal data subsets, the clustering analysis model and the LSTM model are adopted for temporal feature mining and anomaly detection; the output results of each model are weighted and fused to obtain preliminary analysis results, and the weights are dynamically optimized through the particle swarm optimization algorithm.

Preferably, in step S5, the application feedback data includes accuracy evaluation data of analysis results and user demand matching degree data; the performance data includes data collection delay, encryption processing time, classification time, and analysis time; a loss function is constructed based on the feedback data, and the cleaning threshold in the preprocessing stage, the feature weights of the classification model, and the fusion weights of the AI analysis model are adjusted through the gradient descent algorithm to realize closed-loop optimization of processing and analysis strategies.

Preferably, in step S1, the data cleaning algorithm adopts an adaptive threshold filtering algorithm, dynamically setting abnormal value thresholds by statistical data distribution characteristics, and automatically identifying and removing abnormal data exceeding the thresholds; format standardization converts data of different formats (JSON, XML, CSV) into a unified Parquet columnar storage format through a preset multi-source data schema mapping table, improving subsequent processing efficiency.

Preferably, in step S2, the secure transmission protocol adopts a custom transmission protocol based on TLS1.3, adding data integrity check codes during transmission, and verifying whether data is tampered with during transmission through a hash algorithm; after transmission is completed, the receiving end decrypts to obtain the AES key through the RSA private key, and then decrypts to obtain the standardized real-time data stream through the AES key.

Preferably, in step S3, the specific process of multi-dimensional intelligent classification is: inputting the encrypted and desensitized real-time data into the CNN-LSTM fusion model based on the attention mechanism, extracting local features, temporal features, and semantic features of the data; calculating the importance weights of each feature through the attention enhancement layer, and strengthening high-importance features; outputting category labels of the data under each classification dimension by adopting the softmax activation function, generating multi-category data subsets and establishing classification indexes.

Preferably, in step S4, the AI real-time analysis stage further comprises a real-time monitoring module for monitoring the operation status and analysis progress of each AI analysis model, and automatically switching to a standby analysis model when model stalling or analysis delay occurs, ensuring the real-time performance and continuity of analysis.

Beneficial effects

The present invention provides an AI big data real-time processing and analysis method.

Compared with the prior art, the present invention has the following beneficial effects:

The AI big data real-time processing and analysis method adopts a distributed collection architecture (edge nodes + aggregation nodes + central nodes) combined with the stream processing framework Flink to realize parallel collection and real-time circulation of multi-source data; data standardization adopts the Parquet columnar storage format to improve processing efficiency; multi-model parallel analysis and a real-time monitoring fault-tolerance mechanism ensure the continuity of analysis, significantly reducing the delay in data processing and analysis and improving real-time analysis efficiency. It uses the particle swarm optimization algorithm to optimize multi-model fusion weights, fully exploring the local features, temporal features, and deep correlations of data, thus significantly improving the accuracy of analysis results. By matching corresponding analysis models for different types of data, it achieves comprehensive coverage of multi-source heterogeneous data and effectively expands the analysis scope. It adopts a "symmetric encryption + asymmetric encryption" hybrid encryption algorithm to realize full-process encrypted transmission and processing of data, and combines dynamic mask desensitization and differential privacy technology to provide dual protection for user-sensitive information; it prevents data tampering through data integrity verification, realizing full-process security protection for data processing and effectively protecting user privacy. It can automatically extract multi-dimensional features of data and realize accurate classification; combined with a feedback optimization mechanism, it dynamically adjusts model parameters to adapt to the dynamic changes of data features, significantly improving the intelligence and accuracy of data classification.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flow chart of the method of the present invention.

DETAILED DESCRIPTION

The technical solutions in the embodiments of the present invention will be clearly and completely described below in conjunction with the accompanying drawings in the embodiments of the present invention. Obviously, the described embodiments are only a part of the embodiments of the present invention, not all of them. Based on the embodiments of the present invention, all other embodiments obtained by those of ordinary skill in the art without creative work shall fall within the protection scope of the present invention.

Referring to Figure 1, the present invention provides the following technical solutions: an

AI big data real-time processing and analysis method includes the following steps:

S1: real-time data collection and preprocessing stage

This stage adopts a distributed collection architecture to realize real-time collection and preprocessing of multi-source heterogeneous big data, solving the problems of low collection efficiency and high delay of the traditional centralized architecture. The distributed collection architecture comprises edge collection nodes, aggregation nodes, and a central processing node, and each node works collaboratively to realize efficient circulation and standardized processing of data.

The specific process is as follows:

S11: Real-time collection of multi-source data. Edge collection nodes are deployed at various data sources (such as financial transaction systems, government affairs data platforms, industrial sensor networks), adopting the stream processing framework Flink to realize real-time collection and preliminary filtering of data, filtering out obviously invalid data (such as empty data, data with disordered formats); edge collection nodes support the collection of multiple types of data, including structured data (database tables, CSV files), unstructured data (text reports, monitoring images, equipment audio), and temporal data (real-time monitoring data of sensors, temporal data of user behavior).

S12: Data aggregation and format conversion. The aggregation node receives data transmitted by multiple edge collection nodes through a message queue (such as Kafka), aggregates and integrates data from different sources, and performs preliminary format conversion to convert unstructured data into semi-structured data (such as converting text data into JSON format).

S13: Data cleaning and standardization. The central processing node adopts an adaptive threshold filtering algorithm to clean the aggregated data, dynamically setting abnormal value thresholds by statistical data distribution characteristics (mean, variance, median), and automatically identifying and removing duplicate data, abnormal data exceeding the thresholds, and missing value data; through a preset multi-source data schema mapping table, data of different formats (JSON, XML, CSV) are uniformly converted into the Parquet columnar storage format, which has high compression ratio and efficient query performance, and can improve the efficiency of subsequent encryption processing and analysis; finally, a standardized real-time data stream is generated.

S2: Intelligent encryption and secure transmission stage

This stage adopts a "symmetric encryption + asymmetric encryption" hybrid encryption algorithm combined with desensitization technology to realize security protection during data transmission and processing, protect user privacy, and solve the problems of insufficient data security and easy privacy leakage of existing methods.

The specific process is as follows:

S21: Hybrid encryption processing. The AES-256 symmetric encryption algorithm is adopted to batch encrypt the standardized real-time data stream. The AES-256 algorithm has the characteristics of high encryption efficiency and strong security, and can meet the encryption needs of massive real-time data; the RSA-2048 asymmetric encryption algorithm is adopted to encrypt the key of AES-256 to generate a key ciphertext, avoiding the key being stolen during transmission; the encrypted data and the key ciphertext are bound to form an encrypted data unit.

S22: Desensitization of sensitive data. Desensitization is performed on user-sensitive information in the standardized real-time data stream before encryption, adopting a combination of dynamic mask desensitization and differential privacy: retaining the first 6 digits and the last 4 digits of the user's ID number, and replacing the middle digits with "*"; *retaining the first 3 digits and the last 4 digits of the user's mobile phone number, and replacing the middle 4 digits with "*"; retaining the first 6 digits and the last 4 digits of the bank card number, and replacing the middle digits with "*"; adding noise conforming to Gaussian distribution to user behavior data, transaction data, etc., to prevent attackers from deducing user real information through data while ensuring data availability.

S23: Secure transmission. A custom transmission protocol based on TLS1.3 is adopted to transmit the encrypted data unit to the analysis node. The TLS1.3 protocol has the characteristics of fast handshake speed and high security, and can improve transmission efficiency and ensure transmission security; a data integrity check code is added to each encrypted data unit during transmission, and whether the data is tampered with during transmission is verified through the SHA-256 hash algorithm; after transmission is completed, the analysis node decrypts to obtain the AES key through the RSA private key, and then decrypts to obtain the desensitized standardized real-time data stream through the AES key.

S3: Multi-dimensional intelligent classification stage

This stage realizes multi-dimensional intelligent classification of big data based on an

improved deep learning classification model, solving the problems of low intelligence and poor classification accuracy of traditional rule-based classification methods, and laying a foundation for subsequent accurate analysis.

The improved deep learning classification model is a CNN-LSTM fusion model based on an attention mechanism. The model combines the advantages of convolutional neural networks (CNN) in extracting local features and long short-term memory networks (LSTM) in extracting temporal features, and strengthens the influence of key features by introducing an attention mechanism to improve classification accuracy.

The specific process is as follows:

S31: Feature extraction. The encrypted and desensitized standardized real-time data stream is input into the CNN-LSTM fusion model based on the attention mechanism. The feature extraction layer of the model extracts local features of data (such as word vector features of text data, texture features of image data) through CNN convolution kernels, and extracts temporal features of data (such as trend features of temporal monitoring data, sequence features of user behavior) through LSTM memory units; at the same time, preprocessing is performed on unstructured data (word segmentation and word embedding for text data, normalization and convolutional layer feature extraction for image data) to ensure that the data can adapt to model input.

S32: Attention enhancement. The attention enhancement layer assigns weights to the extracted local features and temporal features, and assigns higher weights to key features with high correlation by calculating the correlation between features and classification tasks, strengthening the influence of key features on classification results and suppressing the interference of irrelevant features.

S33: Multi-dimensional classification output. The classification output layer adopts a softmax activation function to realize multi-dimensional classification of data. The classification dimensions include: data source dimension (financial data, government affairs data, industrial data, etc.), data type dimension (structured data, unstructured data, temporal data, etc.), data importance dimension (core data, important data, general data, etc.), and user demand dimension (risk analysis data, trend prediction data, user portrait data, etc.); output category labels of the data under each classification dimension, generate multi-category data subsets and establish classification indexes to facilitate quick calling of subsequent analysis models.

S4: AI real-time analysis stage

This stage constructs a multi-model fusion analysis engine, matching corresponding AI analysis models for different category data subsets to realize real-time in-depth analysis of data, solving the problems of low analysis accuracy and limited analysis scope of existing methods.

The specific process is as follows:

S41: Analysis model matching. Based on the category labels generated by multi-dimensional intelligent classification, corresponding AI analysis models are matched for different category data subsets: for structured data subsets (such as financial transaction data, government affairs statistical data), regression analysis models (such as linear regression, logistic regression) are adopted for trend prediction, and association rule mining models (such as Apriori algorithm) are adopted for data correlation analysis; for unstructured data subsets (such as text reports, monitoring images, equipment audio), deep learning prediction models (such as BERT model for text semantic understanding, CNN model for image target detection, CNN-LSTM model for audio emotion analysis) are adopted for content analysis and semantic understanding; for temporal data subsets (such as real-time monitoring data of industrial sensors, temporal data of user behavior), clustering analysis models (such as K-Means algorithm) are adopted for data clustering, and LSTM models are adopted for temporal feature mining and anomaly detection.

S42: Multi-model fusion analysis. Each matched AI analysis model processes the corresponding data subset in parallel and outputs respective analysis results; the particle swarm optimization algorithm is used to dynamically optimize the fusion weights of the output results of each model, and the output results of each model are weighted and fused to obtain preliminary analysis results; the particle swarm optimization algorithm finds the optimal weight combination through iterative calculation to ensure that the fused analysis results have higher accuracy.

S43: Real-time monitoring and fault-tolerance processing. A real-time monitoring module is set to monitor the operation status (CPU occupancy rate, memory usage rate) and analysis progress of each AI analysis model, and preset delay thresholds and stalling thresholds; when a certain AI analysis model stalls or the analysis delay exceeds the threshold, the real-time monitoring module automatically switches to a standby analysis model (a redundant model with the same function as the original model and optimized parameters) to ensure the real-time performance and continuity of analysis; at the same time, relevant information of the faulty model is recorded for subsequent model optimization.

S5: Result feedback and optimization stage

This stage dynamically optimizes processing and analysis strategies by collecting feedback data, forming a closed-loop optimization mechanism, and continuously improving the efficiency and accuracy of processing and analysis.

The specific process is as follows:

S51: Feedback data collection. Two types of feedback data are collected: one is application feedback data, including accuracy evaluation data of analysis results (obtained through manual review or comparison with actual business results) and user demand matching degree data (obtained through user research or business system feedback); the other is performance data, including data collection delay, encryption processing time, classification time, and analysis time.

S52: Optimization parameter calculation. A loss function is constructed based on the feedback data. The input of the loss function includes analysis result accuracy, user demand matching degree, and total data processing delay. The loss function is minimized through the gradient descent algorithm to calculate the parameters that need to be adjusted, including the cleaning threshold in the preprocessing stage, the feature weights of the classification model, and the fusion weights of the AI analysis model.

S53: Dynamic strategy adjustment. According to the calculated optimization parameters, dynamically adjust the threshold of the adaptive threshold filtering algorithm in the data preprocessing stage, the feature weights of the CNN-LSTM fusion model based on the attention mechanism, and the fusion weights of each model in the multi-model fusion analysis engine; at the same time, optimize the node resource allocation of the distributed collection architecture according to the performance data, and allocate more resources to edge collection nodes or analysis nodes with high load to further improve real-time processing and analysis efficiency. Through continuous feedback and optimization, closed-loop iteration of processing and analysis strategies is realized, and overall performance is continuously improved.

At the same time, the content not described in detail in this specification belongs to the prior art known to those skilled in the art.

It should be noted that in this document, relational terms such as first and second are only used to distinguish one entity or operation from another entity or operation, and do not necessarily require or imply that there is any such actual relationship or order between these

entities or operations. Moreover, the terms "comprise", "include" or any other variants thereof are intended to cover non-exclusive inclusion, so that a process, method, article or equipment including a series of elements not only includes those elements, but also includes other elements not explicitly listed, or elements inherent to such process, method, article or equipment.

Although the embodiments of the present invention have been shown and described, for those skilled in the art, it can be understood that various changes, modifications, substitutions and variations can be made to these embodiments without departing from the principle and spirit of the present invention, and the scope of the present invention is defined by the appended claims and equivalents thereof.