

CLAIMS

1. A method for determination of anomalous data, the method comprising:

retrieving, by one or more processors, a plurality of data packets within a predetermined time period, the plurality of data packets comprising a data structure;

determining, by the one or more processors, a historical distribution, wherein the historical distribution is of historical data comprising the data structure as the data packets;

determining, by the one or more processors, a data distribution state of the plurality of data packets as a current distribution;

determining, by the one or more processors, a first distribution state parameter of a randomly selected data packet in the current distribution;

determining, by the one or more processors, a second distribution state parameter of the randomly selected data packet in the historical distribution;

comparing, by the one or more processors, the plurality of data packets with the historical distribution by determining a difference value between the first distribution state parameter and the second distribution state parameter to generate a comparison result;

determining, by the one or more processors, that the difference value exceeds a predetermined difference threshold;

in response to determining that the difference value exceeds the predetermined difference threshold determining, by the one or more processors, that a data anomaly exists in the plurality of data packets; and

in response to determining that the data anomaly exists in the plurality of data packets, generating, by the one or more processors, an alert indicating the data anomaly.

2. The method according to claim 1, wherein comparing the plurality of data packets with the historical distribution comprises:

generating a plurality of distribution state parameters by substituting the plurality of data packets into the historical distribution; and

comparing the plurality of distribution state parameters with a predetermined threshold related to a distribution state to determine a number of data packets having a distribution state parameter exceeding the predetermined threshold.

3. The method according to claim 1, wherein comparing the plurality of data packets with the historical distribution comprises:

comparing the current distribution with the historical distribution.

4. The method according to claim 3, wherein comparing the current distribution with the historical distribution comprises:

determining a distribution center of the current distribution;

determining a distribution center of the historical distribution;

determining an offset between the distribution center of the current distribution and the distribution center of the historical distribution;

determining that the offset between the distribution center of the current distribution and the distribution center of the historical distribution exceeds a predetermined offset threshold; and

in response to determining that the offset exceeds the predetermined offset threshold, determining that the data anomaly exists in the plurality of data packets.

5. The method according to claim 1, wherein the historical distribution comprises a historical probability distribution determined by processing the historical data with a hybrid Gaussian model.

6. The method according to claim 2, wherein the historical distribution is a historical probability distribution obtained by processing the historical data with a hybrid Gaussian model, and wherein comparing the plurality of data packets with the historical distribution comprises:

substituting the plurality of data packets into the historical probability distribution, thereby obtaining a probability value of each of the plurality of data packets in the historical probability distribution, to obtain a plurality of probability values;

comparing the plurality of probability values with a predetermined probability threshold, and determining a number of the plurality of data packets having a probability value smaller than the predetermined probability threshold, denoted by a first number; and

wherein determining whether data anomaly exists according to the comparison result comprises: determining that data anomaly exists in response to determining that the first number

exceeds a predetermined number threshold, or that a ratio between the first number and the number of the plurality of data packets exceeds a predetermined ratio threshold.

7. The method according to claim 1, wherein the historical distribution is a historical probability distribution obtained by processing the historical data with a hybrid Gaussian model;

determining a data distribution state of the plurality of data packets as a current distribution comprises: processing the plurality of data packets with a hybrid Gaussian model to obtain a current probability distribution; and

comparing the current distribution with the historical distribution comprises comparing the current probability distribution with the historical probability distribution.

8. The method of claim 7, wherein comparing the current probability distribution with the historical probability distribution comprises:

determining a first peak position, where the first peak position is a peak position of a curve corresponding to the current probability distribution;

obtaining a second peak position, where the second peak position is a peak position of a curve corresponding to the historical probability distribution;

determining a position offset between the first peak position and the second peak position; and

wherein determining whether data anomaly exists according to the comparison result comprises: determining that data anomaly exists in response to determining that the position offset exceeds a predetermined offset threshold.

9. The method according to claim 7, wherein comparing the current probability distribution with the historical probability distribution comprises:

determining an occurrence probability of a randomly selected data packet in the current probability distribution, denoted as a first probability;

determining an occurrence probability of the randomly selected data packet in the historical probability distribution, denoted as a second probability;

determining a probability difference value between the first probability and the second probability; and wherein

determining whether data anomaly exists according to the comparison result comprises determining that data anomaly exists in response to determining that the probability difference value exceeds a predetermined difference threshold.

10. The method according to claim 2, wherein the historical distribution is a historical clustering distribution obtained with a clustering algorithm for the historical data, wherein comparing the plurality of data packets with the historical distribution comprises:

determining a corresponding position of each of the plurality of data packets in a distribution space of the historical clustering distribution, to obtain a plurality of positions;

determining a distance between each of the plurality of positions and a corresponding cluster center position of the historical clustering distribution, to obtain a plurality of distances;

comparing the plurality of distances with a preset distance threshold, to determine a number of distances exceeding the preset distance threshold, denoted as a second number; and

wherein determining whether data anomaly exists according to the comparison result comprises: determining that data anomaly exists in response to determining that the second number exceeds a predetermined number threshold, or that a ratio between the second number and the number of the plurality of data packets exceeds a predetermined ratio threshold.

11. The method according to claim 1, wherein the historical distribution is a historical clustering distribution obtained with a clustering algorithm for the historical data, wherein determining a data distribution state of the plurality of data packets as a current distribution comprises:

processing the plurality of data packets with the clustering algorithm to obtain a current clustering distribution; and

wherein comparing the current distribution with the historical distribution comprises: comparing the current clustering distribution with the historical clustering distribution.

12. The method according to claim 11, wherein comparing the current clustering distribution with the historical clustering distribution comprises:

determining a first center position, where the first center position is a center position of a cluster corresponding to the current clustering distribution;

obtaining a second center position, where the second center position is a center position of a cluster corresponding to the historical clustering distribution;

determining a distance between the first center position and the second center position; and

wherein determining whether data anomaly exists according to the comparison result comprises determining that data anomaly exists in response to determining that the distance exceeds a predetermined distance threshold.

13. The method according to claim 11, wherein comparing the current clustering distribution with the historical clustering distribution comprises:

determining a distance between a position of a randomly selected data packet in the current clustering distribution and a center of the cluster corresponding to the current clustering distribution, denoted as a first distance;

determining a distance between a position of the randomly selected data packet in the historical clustering distribution and the center of the cluster corresponding to the current clustering distribution, denoted as a second distance;

determining a distance difference value between the first distance and the second distance; and

wherein determining whether data anomaly exists according to the comparison result comprises determining that data anomaly exists in response to determining that the distance difference value exceeds a predetermined difference threshold.

14. A device for determination of anomalous data, the device comprising a plurality of units and/or modules configured to perform the method of any one of claims 1-13.

15. A computer-readable storage medium having stored thereon a computer program that, when executed on a computer, causes the computer to perform the method of any one of claims 1-13.

16. A computing device, comprising a memory and a processor, wherein the memory stores executable codes, and when the processor executes the executable codes, the processor implements the method according to any one of claims 1-13.