IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 2, 2019, accepted September 17, 2019, date of publication September 25, 2019,
date of current version October 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943817

Medical Data Stream Distribution Pattern Association Rule Mining Algorithm Based on Density Estimation

XIAOFENG LI[1], YANWEI WANG[2], AND DONG LI[3]
[1]Department of Information Engineering, Heilongjiang International University, Harbin 150025, China
[2]Department of Mechanical Engineering, Purdue University, West Lafayette, IN 47906, USA
[3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Corresponding author: Xiaofeng Li (mberse@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61803117, and in part by the Ministry of Education Science and Technology Development Center Industry-University Research Innovation Fund under Grant 2018A01002.

ABSTRACT The traditional data mining method is featured by no analysis over the data distribution and incomplete derived association rule. As a result, the data mining results have the deficiencies of large redundancy probability, large root-mean-square error of approximation (RMSEA) and long consumption time. To handle these issues, this paper proposes a medical data stream distribution pattern association rule mining algorithm based on density estimation. This paper collects medical data, selects the distance method to detect abnormal orphan data in the data stream, detects the duplicate data in the data stream by the similar field matching degree, and eliminates the abnormal data and the duplicate data. Then, the data stream density is estimated based on the histogram estimation samples. According to the data density estimation results, this paper analyzes the distribution of medical data stream from perspectives of concentration, dispersion and morphological characteristics of data distribution. Afterwards, the data distribution pattern association rule mining model is constructed based on compound neural network, data distribution parameters are entered into model's clustering layer, and in-depth training is conducted over the BP (Back Propagation) neural network at the model's mining layer. Meanwhile, all rules under the combination of hidden layer's neuron activity value and corresponding output value, and all rules under the combination of hidden layer's neuron activity value and corresponding input value are derived, so as to complete association rule mining of medical data stream distribution pattern. The experimental results show that the proposed algorithm has a contour curve closest to the true probability density curve; the dispersion degree of medical data is within a reasonable range, and the medical data has high stability; the data redundancy probability is smaller; the mining result's RMSEA is small; data mining takes less time.

INDEX TERMS Compound neural network, medical data, density estimation, distribution pattern, association rules, mining.

I. INTRODUCTION

In recent years, the medical field has witnessed dramatic changes, and the medical data generated has increased exponentially. Due to the increase and enhancement of medical measuring instruments and technologies, large-scale medical information can be accurately recorded and counted, resulting in more sources of medical information data [1], [2]. In particular, as medical information is continuously put into use and used by major medical institutions, the collected patient information includes not only medical images and various physiological indicators, but also a large amount of resources, such as the patients' age and past medical history [3], [4]. When the information and data collected by major medical institutions are gathered together, the summary results are huge and all are the real information and data of patients [5], [6]. The association between them is identified from the above data stream set, and the association rules of its distribution pattern are summarized, which is very beneficial to the treatment of diseases and medical diagnosis [7], [9].

Generally, the mining of the distribution pattern association rule requires the following steps: First, collecting the

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Li.

original data, which often takes a long time; second, identifying the basic distribution conditions by establishing a data distribution histogram; third, conducting the parameter estimation to reflect the medical data distribution characteristics; fourth, performing the fitness test on the collected data [10], [11]. If the test results do not match the assumed distribution, return to the second step for reanalysis [12], [13]. Evidently, the traditional data distribution pattern association rule mining method has the deficiencies of large computation and cumbersome steps [14]. The neural network modeling method is also a good data mining method. The traditional neural network is widely applied, because it has better adaptive learning ability and information processing capability, and provides good data ideas for computer data identification, control and modeling analysis [15], [16]. This traditional neural network method simplifies the modeling process of time series and is of great help to the system research which is difficult to establish mathematical model. Since then, the research on the neural network has attracted the attention of many researchers, and researchers continue to analyze and optimize the traditional neural network.

Foreign researchers have conducted correlation analysis on neural network and data flow distribution. Xing *et al.* [17] self-organizing incremental neural network based on local distributed learning analysis combines the advantages of incremental learning and matrix learning. This method can automatically find suitable nodes and fit learning data incrementally without prior knowledge. Dash and Rao [18] detected arrhythmia based on Wigner-Ville distribution neural network, using Wigner-Ville time-frequency energy distribution and neural network to classify and mine different electrocardiogram beats, the mining accuracy is high. Lal *et al.* [19] used uncertain medical data flowed to estimate the non-invasive blood flow characteristics of cerebral artery, enabled magnetic resonance angiography data to extract specific blood flow rates of patients, and then completed data mining to achieve feature estimation. Su *et al.* [20] automated and efficiently test data streams from different angles, and deeply analyze data shape and distribution problems, which provides a good foundation for data mining. Good results have been achieved through the above research methods, but there are many problems at the same time, and the obtained results have higher redundancy probability.

For medical data mining, a lot of researches have been carried out in China, and numerous research results have been obtained. Domestic scholars take the large-scale data mining in the medical industry as the research object, apply the Apriori algorithm to medical data collection and mining, and use the Apriori algorithm to perform correlation analysis based on the obstetric data in the current medical data. The correlation between cesarean section, physical signs and drugs was discussed, and the relationship between maternal admission time and the number and the distribution of newborns [21]. In order to solve the ECG data processing problem in medical data, based on the Spark cloud platform, an improved genetic K-means clustering algorithm is proposed to be applied to ECG data mining. In this paper, Mallat wavelet transform preprocessing technology is introduced to detect and extract R waves from large-scale ECG data, MIT-BIH data is collected and processed, so as to realize medical ECG data mining [22]. Under the rapid development of the Internet, medical data is characterized by many types and complex relationship. It is difficult for general data processing methods to analyze it effectively. According to the current research status, the origin and characteristics of medical big data are analyzed, and the definition and research status of medical data visualization are introduced. The current medical data visualization mode is converted into two categories, and the cluster describes the most common medical data visualization mining method. The current visualization mode of medical data is converted into two categories, and describes the most common visualization mining method of medical data in a clustering way. The complexity and dynamics have posed challenges to the development of medical data [23]. The two-layer modeling concept under the open EHR specification enhances the flexibility of the medical data system and adapts to the changes in current medical data requirements to some extent. However, due to the large amount of retrieved data, complex query conditions and increasing personalized demands, the retrieval performance of medical data system is relatively weak by using open HER(electronic Human Resource) modeling, which is mainly related to the storage model of underlying data. In order to solve the above problems, based on the two-layer modeling concept under the open EHR specification, this study presents a multidimensional data model mapping method for template and medical dynamic data warehouse, and then improves the implementation speed of ETL(Extract-Load-Transform) through the mapping path to realize medical data stream pattern mining [24].

Medical data stream distribution pattern association rule mining is one of the key ways to understand and identify medical data. The above related research results are not very comprehensive for the cleaning of abnormal data and duplicate data, which leads to higher redundancy probability in the medical data mining process, and large overall mining RMSEA. As a result, this paper proposes a medical data stream distribution pattern association rule mining algorithm based on density estimation. This paper uses the distance method to detect abnormal orphan data in the medical data stream. By estimating the density function of data stream, the data distribution of the data stream can be effectively grasped. Then, the data distribution is analyzed from perspectives of concentration, dispersion and association rules. The association rules are derived from the active values of neurons in the hidden layer, and the clustering group data patterns are mined according to all the rules exported, so as to complete the research task. At the same time, this study uses different experimental indicators to verify the feasibility of the proposed algorithm, providing a good data foundation for the development of medical technology.

The compound neural network is used in the research process. The compound neural network is a compound network structure formed by the combination of neural network learning vector quantization (LVQ) and support vector machine (SVM). The compound neural network framework is usually composed of a standard three-layer feed forward neural network: a data input layer, a hidden layer and a data output layer. The neurons between each layer are weighted and connected, and there is no connection between the neurons in the layer. The compound neural network not only has the advantages of simple structure and high classification efficiency, but also can effectively control the problems caused by the wrong classification at the root and the data deviation, hence the characteristics of strong stability and easy training

Based on the existing research, this paper proposes a medical data stream distribution pattern association rule mining algorithm based on density estimation, and verifies the feasibility of the proposed algorithm. The main research contributions of this paper are:

(1) Histogram density estimation of medical data is conducted, and research on data distribution is carried out, thereby effectively grasping the data distribution conditions and effectively improving the validity of data mining.

(2) Export all the association rules of medical data, and combine the association rule with cluster group data to mine the data.

(3) The validity of the research is verified from many perspectives, so the experimental results are feasible.

II. PRELIMINARY OPERATIONS FOR MEDICAL DATA MINING

The medical data stream often contains some abnormal values. These abnormal values are generally referred to as isolated points [25]. The isolated point is not a random deviation, but is produced in a completely different mechanism. At present, there are many kinds of isolated point detection methods, which can be roughly divided into several types based on statistics, distance and deviation, and each method gives the corresponding definitions of isolated points [26]–[28]. Here, the distance method is selected to implement the abnormal orphan data detection in the medical data stream [29]–[31].

When the abnormal orphan data detection process is performed, the data source containing the abnormal data is obtained for the first time, and the abnormal point in the data source is detected. The isolated point also includes incomplete data points. If there is an isolated point, the database connection cannot be closed, and the isolated point should be removed. If there is no isolated point, the output file is not closed, and the data should be further processed according to the corresponding rules to complete the abnormal point detection [32]–[34].

In order to effectively reduce the redundancy probability of data mining, duplicate data in data stream is detected. The detailed process is as follows:

Assuming that the medical data set $X = \{x_1, \ldots, x_n\}$, the field vector, $F = \{F_1, \ldots, F_p\}$, where F_k represents the p field in the data $x_i = \{x_{i1}, \ldots, x_{ip}\}$, $1 \leq \mathrm{i} < n$. X_{ip} represents the p-dimensional value of the record, and the importance of the field in the target object is represented by the first value, which can also be called attribute weight value. The vector expression of the weight value is

$$W = \{W_1, \ldots, W_p\} \tag{1}$$

According to the above parameter settings, assume that T_{ik} represents the level specified by F_{k} and 1 is the highest level, the higher the value, the lower the level [35]–[37]. T_k represents the ultimate unified level of the k field $k \in \{1, 2, \ldots p\}$, then the expression of T_k is

$$T_k = \left[\frac{\sum_{i=1}^{N} T_{ik}}{N} \right] \tag{2}$$

Assuming that T_k represents the final unified level of F_{k}. T represent the lowest, i.e. the maximum level, and that the final unified level of any two fields is not the same, then $T = p$, using RC transformation to represent the field F_{k} weight value.

$$W_k(RC) = \frac{1}{T} \sum_{t=T}^{T} T_k \frac{1}{t} \tag{3}$$

Assuming that there are two or more fields with the same final unification level, formula (3) can be converted to:

$$W_k = \frac{W_k(RC)}{W^*} \tag{4}$$

where,

$$W^* = \sum_{k=1}^{p} W_k(RC) \tag{5}$$

For any medical data record x_i and x_j, the k-dimensional fields of two are x_{ik} and x_{jk}.The similarity expression of the x_{ik} and x_{jk} fields are as follows:

$$SimField(x_{ik}, x_{jk}) = \frac{\sum_{t=1}^{q} \max(Score(a, x_{ik}))}{|x_{ik}|} \tag{6}$$

where, $Score(a, x_{ik})$ represents the score of the atomic strings a and x_{jk} in x_{ik}, and $|x_{ik}|$ represents the length value x_{ik}, and q represents the number of x_{ik} atomic strings.

To sum up, the expression of duplicate detection for two records is as follows:

$$SimRecord(x_i, x_j) = \sum_{k=1}^{p} SimField(x_{ik}, x_{jk}) W_k \tag{7}$$

According to formula (7), duplicate data in data stream can be detected and eliminated.

III. HISTOGRAM DENSITY ESTIMATION

The medical data stream contains rich and valuable information. Density estimation is a widely used and effective way. If the density function of the data stream is estimated at the time of data input, the data distribution of the data stream can be effectively grasped, and the sparse or dense area can be quickly found in the data stream. Due to data continuous

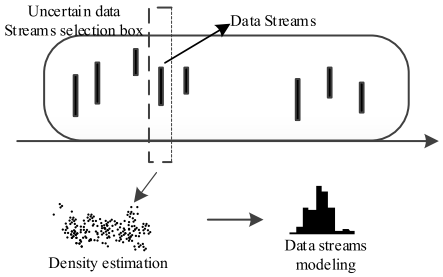


FIGURE 1. Non-parametric modeling method for uncertain data streams.

volume of the medical data stream as well as limited storage space of the data stream and the cache, this paper, without assuming that the samples of the objects conform to any theoretical distribution and based on the data density estimation results, models the medical data by constructing the medical data stream distribution model and selecting the nonparametric estimation method, and then estimates the density of the data stream.

At this time, the density distribution form of medical data mainly depends on the actual value of the sample. The model itself also changes when there is a large change in the sample value. In order to effectively mine the data stream density distribution model, this paper chooses the nonparametric evaluation method of histogram to represent the frequency of medical data and the result is intuitive and easy to understand, which can effectively reduce the complexity of mining calculation. Non-parametric modeling method for uncertain data streams. It is shown in Figure 1.

The key step in histogram density estimation is to determine the class interval d. The nature of the density function corresponding to different class intervals and the overall distribution characteristics also vary. If the class interval d is unreasonable, the estimated density function will have a large deviation from the real situation. Therefore, assuming that the true density function of the uncertain object u is $u(x)$, and histogram density estimation is $\widehat{u}(x)$.Combined with the above theory, the method of histogram density estimation for uncertain objects is given. The steps are as follows:

(1) Suppose the sample of the uncertain data object is $u_{i,j}^I$, the size is represented by I, and the sample is sorted; the minimum and maximum samples are $x_1 = \min\left(u_{i,j}^I\right)$ and $x_n = \max\left(u_{i,j}^I\right)$ respectively;

(2) Estimate the class interval d;

(3) Calculate the class interval $A_i = (a_i, a_{i+1}), i = 0, 1, \ldots, k-1$, and the sample frequency is $v_1, v_2, \ldots, v_k$;

(4) Make a histogram with d as the bandwidth and $v_1, v_2, \ldots, v_k$ as the height;

(5) $\widehat{u}(x)$ can be obtained by estimating the density distribution of the population corresponding to the samples according to the histogram.

IV. MEDICAL DATA DISTRIBUTION PATTERN

According to the statistical principle, the data distribution is mainly composed of the following three characteristics: data concentration, dispersion and morphological characteristics of data distribution.

A. MEDICAL DATA CONCENTRATION

The data concentration refers to the degree to which the data is close to or concentrated toward the center value. The main indicators for determining central tendency can be divided into the following two categories: position representative value and numerical mean. Between them, the position representative value includes the mode, median and quartile, etc [38]–[40]. The numerical mean includes the arithmetic mean, the harmonic mean and geometric mean. There are three quartiles: the lower quartile, the median, and the upper quartile.

B. MEDICAL DATA DISPERSION

Medical data dispersion is a statistical description of data variation. Its main role is to measure representative nature and determine the uniformity and stability of data changes. Dispersion is measured by the following indicators: range, quartile range, mean difference, variance, standard deviation, and dispersion coefficient. The dispersion coefficient is also called the variation coefficient. It is the ratio of the dispersion indicator to the arithmetic mean, representing a dimensionless quantity. It is suitable for analysis of variations in data with different phenomena or levels. The dispersion coefficient is divided into range coefficient, coefficient of average deviation, and coefficient of standard deviation. Among them, the coefficient of standard deviation V is commonly used, which is the ratio of standard deviation σ to arithmetic mean g.

C. MORPHOLOGICAL CHARACTERISTICS OF DATA DISTRIBUTION

The morphological characteristics of data distribution mainly include moment, kurtosis and skewness. Among them, the moment means the arithmetic mean of all variables and the n-th power constant dispersion. This paper uses the third order moment and the fourth order moment to characterize the morphological features of data distribution. Among them, the calculation formula for the third-order center distance v_3 and the fourth-order center distance v_4 is:

$$v_3 = \frac{\sum_{i=1}^{N}\left(x_i' - \bar{x}\right)^3}{n} \tag{8}$$

$$v_4 = \frac{\sum_{i=1}^{N}\left(x_i' - \bar{x}\right)^4}{n} \tag{9}$$

In the above formula, x_i' is the data morphological characteristic value, and $\bar{x}$ is the mean value of x_i'.

Skewness represents the direction and extent of data asymmetry, which can be divided into Pearson skewness, quartile skewness and moment method skewness. Among them, the moment method skewness is more accurate. Normally,

TABLE 1. Distribution pattern association rules subdivision Table.

Distribution type	Distribution function parameter（l_1 is the scale parameter, l_2 is the shape parameter）	Number of segments
index distribution	l_1=1~100，Interval is 1	100
Normal distribution	Average is 0，The standard deviation is between 1 and 100. Interval is 1	100
Lognormal distribution	l_1=0; l_2=0 1~10，Interval is 0.1	100
Weibull distribution	l_1=1; l_2=0 5~50，Interval is 0.5；l_2=1，is index distribution	99
Gamma distribution	l_1=1; l_2=0 1~10，Interval is 0.1；l_2=1，is index distribution	99

the skewness index α represents the cube of the ratio of the third-order center distance to the standard deviation σ, expressed as:

$$\alpha = (v_3/\sigma)^3 \tag{10}$$

Peak is also called kurtosis, indicating the sharpness and peak degree of the medical data distribution pattern. The quantile kurtosis represents a measured value of the kurtosis. According to the specific concept of the moment, the kurtosis is accurately measured to obtain the moment kurtosis. The specific calculation method is as follows:

$$\beta = (v_4/\sigma)^4 \tag{11}$$

This paper combines the above three medical data distribution characteristics to determine the quartile, namely, the lower quartile, the median, the upper quartile, as well as the dispersion coefficient, skewness (α) and kurtosis (β). The medical data distribution pattern is determined by the above six parameters.

V. COMPOUND NEURAL NETWORK DATA DISTRIBUTION PATTERN ASSOCIATION RULE MINING

Based on the morphological characteristics of the above medical data distribution and according to the probability density function shape under different distribution parameters, this paper divides each distribution pattern association rule into different distribution parameters, which is roughly subdivided into 498 distribution pattern association rules. It is shown in table 1.

In practical applications, a large amount of data may obey the probability distribution. This section builds the data distribution pattern association rule mining model. The model is mainly divided into two layers: the first layer is to cluster the data distribution, and the data with similar characteristics is classified into the same class; the second layer is to mine the classified features by using BP neural network.

Usually, multiple experiments are used when constructing network data samples. In this study, data distribution pattern association rule in Table 1 is used as network parameter input, data distribution pattern association rule is used as network data output, and then distribution pattern association rule mining of medical data stream is completed based on compound neural network. Specific steps are as follows:

A. DATA NORMALIZATION PROCESSING

The samples that need to be input in the compound neural network have the same data length. The data samples are normalized in this paper. Assume that the input data sample set is S, and the data reference mode set is E, which are expressed as:

$$S = \{s_1, s_2, \ldots, s_m\} \tag{12}$$

$$E = \{e_1, e_2, \ldots, e_m\} \tag{13}$$

In the above formula, s_m represents the coordinate vector of the m-th point of S, and e_m represents the coordinate vector of the m-th point of E. A unified mapping transformation is performed on the data sample set and the data reference pattern set, and S is converted to a new position vector S', namely:

$$S' = Js_m + bE \tag{14}$$

In the above formula, J represents a two-dimensional matrix, and b is a variation vector coefficient.

The normalized sample is:

$$S' = \{s'_1, s'_2, \ldots, s'_m\} \tag{15}$$

B. INPUT AND OUTPUT OF DATA DISTRIBUTION PATTERN ASSOCIATION RULE

In the first layer of the network layer, this study inputs characteristic parameters such as correlation coefficient, variance, mean, skewness and cumulative probability into the network, makes statistics over the data distribution pattern association rule, and classifies the distribution pattern association rule, so as to obtain different classification groups and memory rights for input sample patterns.

The data obtained by the above classification is trained in the BP neural network, and the final network output is the data distribution pattern association rule. In the BP neural network training, even if there is only one output layer node, as the input unit increases, the number of connections between nodes in the network will continue to increase exponentially, which leads to an increase in the corresponding extraction rules. Besides, this brings a lot of difficulties to the extraction of mining rules for the medical data stream distribution pattern association rules. As a result, the network needs to be further trained to facilitate the extraction of easy-to-understand mining rules [41], [42].

The training process of the neural network is as follows:

Assuming that the input mode of the neural network is x_l^I, $I \in \{1, \ldots, K\}$, $I \in \{1, \ldots, M\}$, l represents the set of input tuples in the data set. K is the maximum number of tuple sets. I is the input mode dimension. M is the maximum mode dimension. Assuming that W_l^m represents the connection weight of the l-th neuron in the input layer to the m-th neuron in the hidden layer. Then the activation value of the m-th neuron in the hidden layer can be expressed as:

$$\partial^m = f(\sum_{l=1}^{M} (x_l^I W_l^m) - R^m) \tag{16}$$

where, $f(\cdot)$ represents the neuron activation function of hidden layer, which is non-linear function. $f(\cdot)$ is defined as hyperbolic tangent function. R^m represents the m-th neuron bias of the hidden layer, which maps the larger value of the activation $[-1, 1]$. If the activation of neuron in the hidden layer is obtained, the output value of the P-th neuron in the output layer can be expressed as:

$$S_P^1 = \sigma(\sum_{m=1}^{h} \partial^m V_P^m) \tag{17}$$

where, $\sigma(\cdot)$ represents the neuron activation function in the output layer, $\sigma(\cdot)$ is defined as a simiod function, which maps the larger value of activation value to a smaller interval [0, 1]. V_P^m represents the connection weight of the m-th neuron in the hidden layer to the P-th neuron in the output layer.

An input tuple can be accurately classified, assume that it satisfies the $\max |S_P^I - t_P^I| \leq \gamma_1$ condition, where γ_1 represents a positive number less than 0.5. Assuming that the input mode x^I belong to C_J, and its target output is expressed as an o-dimensional vector t_P^I, $P \in \{1, 2, \ldots, O\}$. When $P = J$, $t_P^I = 1$, and conversely $t_P^I = 0$.

The main purpose of neural network training is to find a set of optimal connection weights, and then to enable the network to achieve element classification with high accuracy [43]. For the purpose of measuring the classification error value, the training process is transformed into a process of minimizing the error function by adjusting the connection weights (w, v), the error function is cross-entropy function $E(w, v)$. The function can be expressed as:

$$E(w, v) = -\sum_{I=1}^{K} \sum_{P=1}^{o} (t_P^I \log S_P^I + (1 - t_P^I) \log(1 - S_P^I)) \tag{18}$$

The purpose is to convert more connection weights into a very small value, which is convenient for network pruning. $P(w, v)$ is introduced. The expression of the compensation term is as follows:

$$\begin{aligned} P(w, v) = \varepsilon_1 (& \sum_{m=1}^{M} \sum_{l=1}^{M} \frac{\beta (w_1^m)^2}{1 + \beta (w_l^m)^2} \\ & + \sum_{m=1}^{h} \sum_{P=1}^{o} \frac{\beta (v_P^m)^2}{1 + \beta (v_P^m)^2}) \\ & + \varepsilon_2 (\sum_{m=1}^{M} \sum_{l=1}^{M} (W_1^m)^2 \\ & + \sum_{m=1}^{h} \sum_{P=1}^{o} (V_P^m)^2) \end{aligned} \tag{19}$$

where, ε_1 and ε_2 represent two positive weight decay parameters, h represents the number of neurons in the hidden layer. O represents the number of neurons in the output layer.

In the training process of neural network, we should first assign very small random number $(w, v)^{(0)}$ to each weight value, such as $[-1, 1]$, and then iteratively update the weight value, so that $E(w, v) + P(w, v)$ is the smallest and the training is over.

Subsequently, this paper randomly selects any of the cluster groups, and describes the corresponding data distribution (Normal distribution, Beta distribution and Weibull distribution), as well as the corresponding hidden layer's neural network and number of neurons. Among them, the number of neurons is determined by the data distribution pattern association rule. The other classification groups are described in a similar manner.

C. RULE EXTRACTION AND DATA PROBABILITY PATTERN MINING IMPLEMENTATION

In this paper, the random data distribution sequence selected by computer is used to calculate the eigenvalues of the data sequence and input it into the network. The data is clustered in the first layer network layer and belongs to each cluster group. After training the BP neural network at the second layer, the cluster group data pattern is mined according to all the rules that are derived.

Although neural network pruning can simplify the whole structure, it is still difficult to find a clear correlation between input tuples and output tuples. The main reason is that it is more difficult to extract mining rules [44]. According to the requirements of reality and mining accuracy, the steps of extracting mining rules from distribution patterns of medical data streams are set as follows:

(1) Discreteness and classification of active values

Set $\varepsilon \in (0, 1)$, D represents the number of discrete active neurons in the hidden layer, and δ_1 represents the number of active neurons in the hidden layer during the input of the first mode in the training sample set. Suppose $H(1) = \delta_1$, count(1) = 1, sum(1) = δ_1, D = 1.

For all patterns $I = 1, 2, \ldots,$ K, in the training sample set, assuming that δ is its active value. If the subscript $\bar{J}$ can make $|\delta - H(\bar{J})| = \min_{J \in \{1,2,..,D\}} \delta - H(J)$ and $|\delta - H(\bar{J})| < \varepsilon$, then $count(\bar{J}) = count(\bar{J}) + 1$, $sum(D) = sum(D) + \delta$, on the contrary, D = D + 1, $h(D) = \delta$, count(D) = 1, $sum(D) = \delta$.

Substitute H, $H(J) = sum(J)/count(J)$, $J = 1, 2, \ldots, D$, according to the mean of activity value subordinated to this category. Replacing δ^I with active value δ_d of subordinate category of neuron activity value δ^I in hidden layer and its accuracy is tested. Assuming that its accuracy is lower than the required level, the value of ε is reduced and step 1 is performed iteratively.

(2) Enumerating discrete active values and calculating network output. All rules are derived by combining the active value of hidden layer neurons with the corresponding output value.

TABLE 2. Compound neural network structural parameters.

Parameter	value
Dimension	[1-50]
Number of convolutions	2
Number of iterations	60
Weights	[-100,100]

(3) In view of the above active values of neurons in the hidden layer, the corresponding input values are enumerated, and all rules under the combination of the active values and the corresponding input values are derived.

(4) According to the rules generated in (2) and (3), all the rules under the combination of input value and output value are derived, and the final result is the mining rules of probability distribution pattern of medical data stream, according to which the research purpose of the subject can be achieved.

VI. ANALYSIS OF EXPERIMENTAL RESULTS

The experiment was carried out in the environment of VC++6.0 and MATLAB. Considering the advantages and disadvantages of programming conditions, in the process of implementing medical data stream distribution pattern association rule mining algorithm based on density estimation, the algorithm functions are realized by calling the interface functions of MATLAB and VC++6.0 in VC++6.0, with VC++6.0 as the main part and MATLAB as the auxiliary part. In this experiment, 550 groups of medical data from ADNI (the Alzheimer's Disease Neuroimaging Initiative) database (adni. loni. usc. edu) were used to ensure that the experimental data were normal. 550 groups of images were disorganized in order. 150 groups of data were used as abnormal sample data, 400 groups of data were used as test data, and the number of repeated experimental data analysis was up to 110 times.

The structural parameters of the compound neural network are shown in Table 2.

The experimental indicators are as follows:

1. Data redundancy probability

In the early stage of medical data mining, redundant data was eliminated and data quality was improved, in order to verify the performance of this operation, the data redundancy probability is selected as the indicator for analysis. The formula for calculating the data redundancy probability is as follows:

$$R = \frac{C}{G} \times 100\% \tag{20}$$

In the above formula, C represents the amount of repeated medical data, and R represents the total amount of medical data.

2. Histogram density estimation

In this paper, we study the mining of association rules based on density estimation of medical data flow distribution pattern. Density estimation is a novel point in this paper. It is necessary to analyze the operation effect of this step.

3. Stability of medical data distribution pattern

Based on the distribution of medical data sample points, the stability of the method is verified.

4. Mining result approximate root mean square error

The root-mean-square error is the deviation between the true value and the theoretical value. The root-mean-square error is an important indicator to measure the performance of the data distribution association rule. If RMSEA is less than 0.08, it means that the performance of the research results is superior; if the root-mean-square error is larger, it indicates that the research results are poor. The formula for calculating the indicator is as follows:

$$Q = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})}{q}} \tag{21}$$

In the above formula, $\bar{X}$ represents the average value of the medical data set, and q represents the number of experiments.

5. Time-consuming data mining

Taking the data mining time as an indicator, the method is compared with the literature [22], literature [23], and literature [24] to verify the performance of the proposed method.

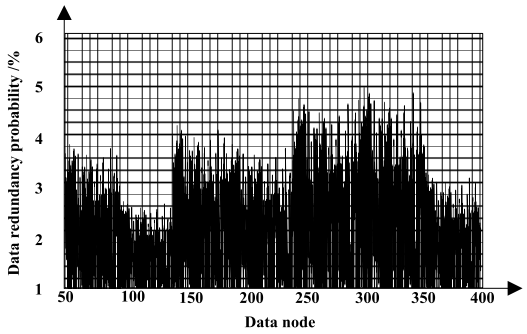
A. DATA REDUNDANCY PROBABILITY COMPARISON

Duplicate data detection is an important step in mining probability distribution patterns of medical data streams. The main purpose is to eliminate redundant data. In order to fully verify the superiority of this algorithm, the probability of redundant data is used as an index for comparative analysis. The results are shown in Figure 2.

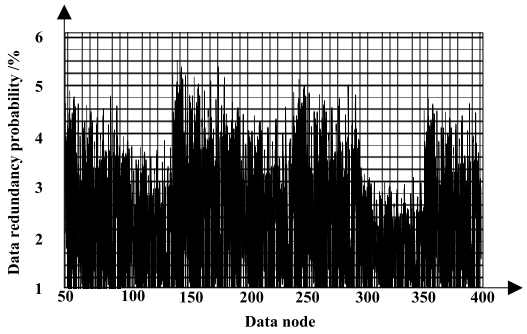
The data nodes in Figure 2 are the components of the medical data stream. The higher the probability of data redundancy is, the more repetitive actions exist in the mining process, which will seriously reduce the mining efficiency and affect the mining accuracy to a certain extent. The lower the probability of data redundancy, the higher the mining efficiency and the lower the energy consumption. As shown in Figure 2, the proposed medical data stream distribution pattern association rule mining algorithm based on density estimation is well controlled under different number of nodes, with the probability of data redundancy not exceeding 2%, which is much lower than the results of literature research. The proposed algorithm detects duplicate data in medical data stream by data field similarity registration, reduces the probability of data redundancy, and effectively enhances the real-time mining.

B. HISTOGRAM DENSITY ESTIMATION

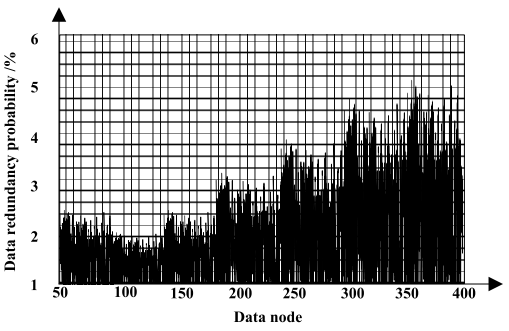
In the data stream distribution pattern association rule mining process, the histogram density estimation reflects the true probability density function $u(x)$ and the probability density estimation $\widehat{u}(x)$. The smaller the degree of advancement is, the closer the upper contour of the histogram is to the parameter probability density function, the better the grouping effect is, and the more accurate histogram density estimation will be; vice versa, the worse the histogram density estimation is.



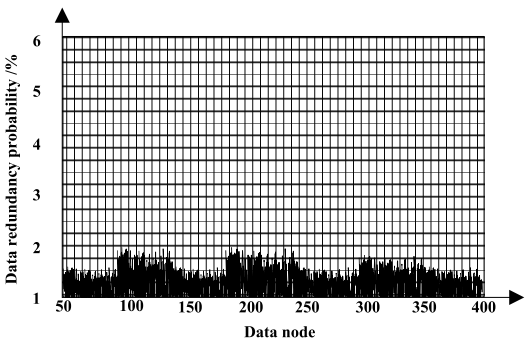
(a) Data redundancy probability of literature [17]



(b) Data redundancy probability of literature [18]



(c) Data redundancy probability of literature [19]



(d) Data redundancy probability of the proposed algorithm

FIGURE 2. Data redundancy probability comparison of different research results.

According to Figure 3, compared with other literature methods, the contour curve of the method has the closest degree to the true probability density curve. The main reason is that this paper constructs the medical data stream distribution model and selects the non-parametric estimation

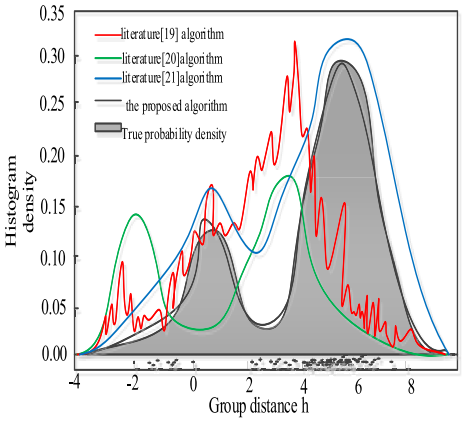


FIGURE 3. Histogram density estimation curve.

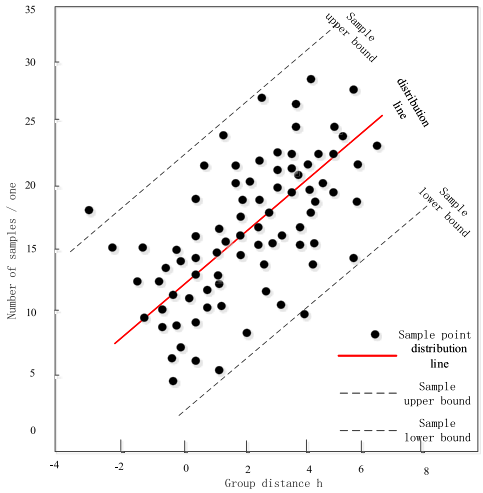


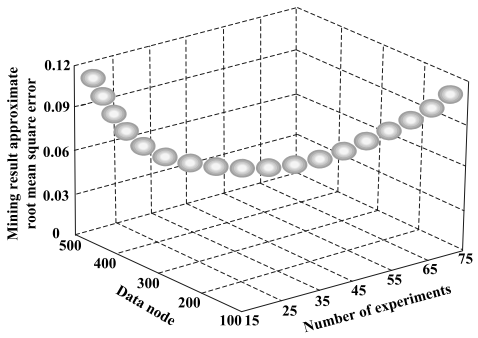
FIGURE 4. Medical data distribution pattern.

method to model the medical data, so it can effectively grasp the data distribution of the data stream and quickly find the sparse or dense areas in the data stream. Therefore, after the proposed algorithm is applied, the histogram density estimation is more accurate, which lays a good foundation for the mining of medical data stream distribution pattern association rule.

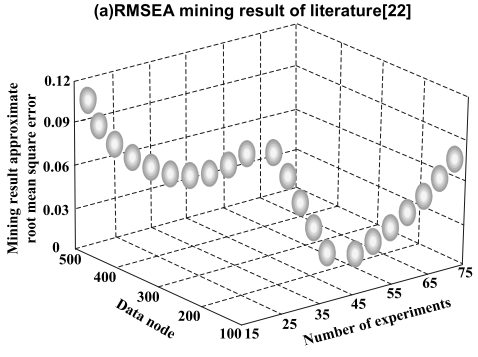
C. STABILITY OF MEDICAL DATA DISTRIBUTION PATTERN

The distribution characteristics of medical data are mainly used to determine the uniformity and stability of medical data changes. The tighter the distribution of sample points, the higher the stability of the medical data distribution, and vice versa.

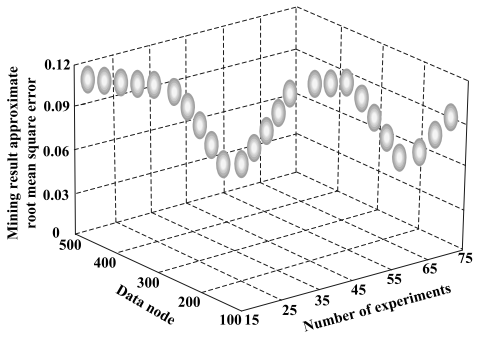
According to Figure 4, the sample points are closely distributed on the distribution line, which proves that the medical data distribution of the proposed algorithm has high stability performance. The main reason is that the proposed algorithm analyzes the three characteristics of data concentration, dispersion and morphological characteristics of data distribution, and obtains the distribution pattern of medical data, thereby having good stability.



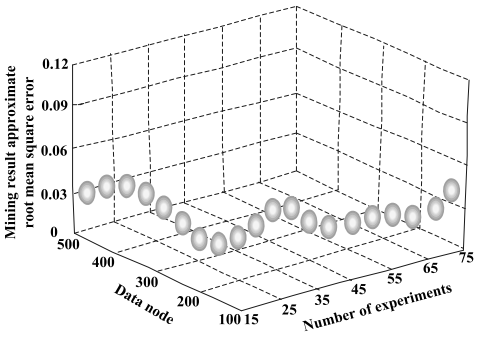
(a)RMSEA mining result of literature[22]



(b)RMSEA mining result of literature[23]



(c)RMSEA mining result of literature[24]



(d)RMSEA mining result of the proposed algorithm

FIGURE 5. Comparison of RMSEA of different literature.

D. COMPARISON OF RMSEA

The research results of this paper are compared with the root-mean-square error of the literature research results. It is shown in Figure 5.

TABLE 3. Time-consuming comparison of mining.

Data volume	Literature[22] /s	Literature[23] /s	Literature[24] /s	Research results in this paper /s
50	23	26	30	10
100	25	25	31	11
150	26	23	36	13
200	30	30	29	9
250	29	28	40	10
300	35	31	36	12
350	36	36	38	16
400	41	39	26	10

According to Figure 5, although several results presented in the literature also have small errors, the overall fluctuation of the error value is relatively large, the reliability is low, and many data points have the RMSEA values exceeding 0.08. The overall RMSEA value of the proposed algorithm mining results is less than 0.03, which shows good mining performance. The result is mainly due to the proposed algorithm pre-processing the medical data before carrying out the excavation operation and training the network after constructing the neural network. Therefore, the proposed algorithm facilitates the extraction of easy-to-understand mining rules, enhances the mining performance, and improves the mining accuracy.

E. TIME-CONSUMING COMPARISON OF MINING

The research results of this paper are compared with mining time-consuming of literature research results. It is shown in Table 3.

Table 3 shows that the mining results of the literature research are time-consuming, which is equivalent to about 3.5 times of the proposed mining algorithm, indicating that the mining of the medical data stream distribution pattern association rule is less time-consuming and can be quickly completed. The reason is that the proposed algorithm uses the active value δ_d of the active value of the neuron in the hidden layer to replace the δ^I, and the accuracy is detected, so it is easier to find a clear correlation between the input tuples and the output tuples, reducing the data mining time.

VII. CONCLUSION

Medical data mining is a feasible way to improve medical technology and patient satisfaction. In order to deal with the high probability of redundancy and large mining result RMSEA in the existing research results, combined with actual needs, this paper proposes a medical data stream distribution pattern association rule mining algorithm based on density estimation. By taking the pre-processing of medical data as the mining support, the author adopts the compound neural network algorithm to construct the neural network, trains the network and extracts the medical data stream distribution pattern association rules, and completes the mining research. Experiments and discussions show that the proposed mining algorithm obtains more accurate histogram density estimation; the medical data has higher stability and

can effectively detect duplicate data and reduce data redundancy; the mining results have a RMSEA of 0.03, with less mining time. Therefore, the proposed algorithm is a feasible medical data mining algorithm. It can effectively solve the problems existing in the existing research results.

The author has done a lot of data preparation works in the research process and achieved good research results. However, in-depth research is needed for adapting to the actual situation.

In this paper, the medical data flow distribution pattern association rule mining is carried out under an idealized condition. If there is a single or even multiple external disturbances in actual operations and applications, it may affect the research results. Therefore, the next step will be to analyze the distribution pattern and density of the data under different external disturbance conditions and deep mining data features, further improve the performance of the mining algorithm and effectively apply it in the medical field.

REFERENCES

[1] A. Villar, M. T. Zarrabeitia, A. Santurtún, and P. Fdez-Arroyabe, "Integrating and analyzing medical and environmental data using ETL and business intelligence tools," *Int. J. Biometeorology*, vol. 62, no. 6, pp. 1085–1095, 2018.

[2] M. D. Kohli, R. M. Summers, and J. R. Geis, "Medical image data and datasets in the era of machine learning—Whitepaper from the 2016 C-MIMI meeting dataset session," *J. Digit. Imag.*, vol. 30, no. 4, pp. 392–399, 2017.

[3] Z. Hass, M. Levine, J. Ting, H. Xu, and L. P. Sands, "The modeling of medical expenditure data from a longitudinal survey using the generalized method of moments (GMM) approach," *Statist. Med.*, vol. 35, no. 15, pp. 2652–2664, 2016.

[4] A. C. Constantinou, N. Fenton, L. Radlinski, and W. Marsh, "From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support," *Artif. Intell. Med.*, vol. 67, pp. 75–93, Feb. 2016.

[5] H. Gao, K. Zhang, J. Yang, F. Wu, and H. Liu, "Applying improved particle swarm optimization for dynamic service composition focusing on quality of service evaluations under hybrid networks," *Int. J. Distrib. Sensor Netw.*, vol. 14, no. 2, pp. 1–14, 2018.

[6] Y. Yin, L. Chen, Y. Xu, and J. Wan, "Location-aware service recommendation with enhanced probabilistic matrix factorization," *IEEE Access*, vol. 6, pp. 62815–62825, 2018.

[7] S. K. Chung, "Relationships between safety perception, knowledge, and compliance among hospital employees," *Asia–Pacific J. Convergent Res. Interchange*, vol. 4, no. 4, pp. 71–80, 2018.

[8] B. N. Harini and T. Rao, "An extensive review on recent emerging applications of artificial intelligence," *Asia–Pacific J. Convergent Res. Interchange*, vol. 5, no. 2, pp. 79–88, 2019.

[9] Y. Yin, L. Chen, Y. Xu, J. Wan, H. Zhang, and Z. Mai, "QoS prediction for service recommendation with deep feature learning in edge computing environment," *Mobile Netw. Appl.*, vol. 4, pp. 1–11, Apr. 2019.

[10] P. Ranjan, N. Dhaka, I. Pant, and A. Pranav, "Design and analysis of two stage op-amp for bio-medical application," *Int. J. Wearable Device*, vol. 3, no. 1, pp. 9–16, 2016.

[11] H.-C. Jeong, "An analysis on nursing Students' perception of protective actions and management for medical information at hospital," *Int. J. Reliable Inf. Assurance*, vol. 4, no. 2, pp. 19–24, 2016.

[12] J.-H. Huh, H. B. Kim, and J. A. Kim, "Method of modeling of basic big data analysis for Korean medical tourism: A machine learning approach using apriori algorithm," in *Information Science and Applications* (Lecture Notes in Electrical Engineering), vol. 424, K. Kim and N. Joukov, Eds. Singapore: Springer, 2017.

[13] J. Gangyong, H. Guangjie, J. Lloret, and L. Aohan, "Coordinate channel-aware page mapping policy and memory scheduling for reducing memory interference among multimedia applications," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2839–2851, Dec. 2017.

[14] X. J. Shi and Y. S. Cai, "A sliding window based method for weighted frequent-pattern mining over data stream," *Intell. Comput. Appl.*, vol. 8, no. 2, pp. 63–67, 2018.

[15] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4014–4024, Dec. 2017.

[16] J. Congfeng, Y. Wang, D. Ou, Y. Li, J. Zhang, J. Wan, B. Luo, and W. Shi, "Energy efficiency comparison of hypervisors," *Sustain. Comput., Inform. Syst.*, vol. 22, pp. 311–321, Jun. 2019.

[17] Y. Xing, X. Shi, K. Zhou, J. Zhao, and F. Shen, "A self-organizing incremental neural network based on local distribution learning," *Neural Netw.*, vol. 84, pp. 143–160, Dec. 2016.

[18] S. K. Dash and G. S. Rao, "Arrhythmia detection using Wigner–Ville distribution based neural network," *Procedia Comput. Sci.*, vol. 85, no. 2, pp. 806–811, 2016.

[19] R. Lal, F. Nicoud, E. Le Bars, J. Deverdun, F. Molino, V. Costalat, and B. Mohammadi, "Non invasive blood flow features estimation in cerebral arteries from uncertain medical data," *Ann. Biomed. Eng.*, vol. 45, no. 11, pp. 2574–2591, 2017.

[20] T. Su, K. Wu, W. Miao, G. Pu, J. He, Y. Chen, and Z. Su, "A survey on data-flow testing," *ACM Comput. Surv.*, vol. 50, no. 1, 2017, Art. no. 5.

[21] K. Jia, H. Li, and Y. Yuan, "Application of data mining in mobile health system based on Apriori algorithm," *J. Beijing Univ. Technol.*, vol. 43, no. 3, pp. 394–401, 2017.

[22] B.-Z. Che, C.-J. Wei, and X.-L. Wan, "Big data analysis and processing of ECG based on Spark platform," *Comput. Eng. Design*, vol. 39, no. 1, pp. 108–114, 2018.

[23] Y. Wang and S. X. Ren, "Survey on visualization of medical big data," *J. Frontiers Comput. Sci. Technol.*, vol. 11, no. 5, pp. 681–699, 2017.

[24] D. D. Wang, F. Ye, and X. D. Lv, "Construction method of clinical medical data warehouse based on openEHR," *Chin. J. Biomed. Eng.*, vol. 35, no. 2, pp. 162–168, 2016.

[25] N. Yaraghi and R. D. Gopal, "The role of HIPAA omnibus rules in reducing the frequency of medical data breaches: Insights from an empirical study," *Milbank Quart.*, vol. 96, no. 1, pp. 144–166, 2018.

[26] M. Amy, "AI researchers embrace Bitcoin technology to share medical data," *Nature*, vol. 555, no. 7696, pp. 293–294, 2018.

[27] H. Gao, D. Chu, Y. Yin, and Y. Duan, "Probabilistic model checking-based service selection method for business process modeling," *J. Softw. Eng. Knowl. Eng.*, vol. 27, no. 6, pp. 897–923, 2017.

[28] P. H. Oliveira, L. C. Scabora, M. T. Cazzolato, W. D. Oliveira, R. S. Paixão, A. J. M. Traina, and C. Traina, "Employing domain indexes to efficiently query medical data from multiple repositories," *IEEE J. Biomed. Health Inform.*, to be published.

[29] X. H. Tian and Z. Y. Chen, "Research on high dimensional data mining technology oriented to large data," *Automat. Instrum.*, vol. 26, no. 3, pp. 100–101, 2016.

[30] H. Gao, Y. Duan, H. Miao, and Y. Yin, "An approach to data consistency checking for the dynamic replacement of service process," *IEEE Access*, vol. 5, pp. 11700–11711, 2017.

[31] X. Y. Zhang and C. N. Fu, "An efficiency mining algorithm for multiple class data," *J. China Acad. Electron. Inf. Technol.*, vol. 12, no. 4, pp. 359–364, 2017.

[32] L. Shen, H. Chen, Y. Zhe, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, "Evolving support vector machines using fruit fly optimization for medical data classification," *Knowl.-Based Syst.*, vol. 96, pp. 61–75, Mar. 2016.

[33] K. A. Padrez, L. Ungar, R. J. Smith, S. Hill, T. Antanavicius, D. M. Brown, P. Crutchley, D. A. Asch, and H. A. Schwartz, "Linking social media and medical record data: A study of adults presenting to an academic, urban emergency department," *BMJ Qual. Saf.*, vol. 25, no. 6, pp. 414–423, Jun. 2016.

[34] L. Gottlieb, R. Tobey, D. Hessler, N. E. Adler, and J. Cantor, "Integrating social and medical data to improve population health: Opportunities and barriers," *Health Affairs*, vol. 35, no. 11, pp. 2116–2123, 2016.

[35] J. L. Bruse, M. A. Zuluaga, A. Khushnood, K. McLeod, H. N. Ntsinjana, T.-Y. Hsia, M. Sermesant, X. Pennec, A. M. Taylor, and S. Schievano, "Detecting clinically meaningful shape clusters in medical image data: Metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 10, pp. 2373–2383, Oct. 2017.

[36] T. Shaikhina and N. A. Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," *Artif. Intell. Med.*, vol. 75, pp. 51–63, Jan. 2017.

[37] X. Zhou, Y. Li, L. Zhao, and J. Yuan, "An improved fuzzy neural network compound control scheme for inertially stabilized platform for aerial remote sensing applications," *Int. J. Aerosp. Eng.*, vol. 2018, Aug. 2018, Art. no. 7021038.
[38] Y. B. Lv, J. W. Zhao, and F. L. Cao, "Image denoising algorithmbased on composite convolutional neural network," *Pattern Recognit. Artif. Intell.*, vol. 30, no. 2, pp. 97–105, 2017.
[39] Z.-M. Bi and Y. Wang, "Method of flatness pattern recognition based on improved genetic algorithm optimization Elman neural network," *J. Iron Steel Res.*, vol. 29, no. 4, pp. 305–311, 2017.
[40] M. A. P. Chamikara, P. Bertok, S. Camtepe, I. Khalil, and D. Liu, "Efficient data perturbation for privacy preserving and accurate data stream mining," *Pervasive Mobile Comput.*, vol. 48, pp. 1–19, Aug. 2018.
[41] J. Song and X. Chen, "Weighted frequent pattern tree algorithm of transform data stream with time," *J. Jiangsu Univ. (Natural Sci. Ed.)*, vol. 38, no. 3, pp. 330–335, 2017.
[42] A. Altomare, E. Cesario, C. Comito, F. Marozzo, and D. Talia, "Trajectory pattern mining for urban computing in the cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 2, pp. 586–599, Feb. 2017.
[43] C. Yu, Y. Li, M. Zhang, and H. Xiang, "Data mining-assisted short-term wind speed forecasting by wavelet packet decomposition and Elman neural network," *J. Wind Eng. Ind. Aerodynamics*, vol. 175, pp. 136–143, Jan. 2018.
[44] Z. Ning, D. C. Shao, and Y. Chen, "Android malware detection system based on signature and data flow pattern mining," *Comput. Sci.*, vol. 44, no. z2, pp. 327–331, 2017.

XIAOFENG LI received the Ph.D. degree from the Beijing Institute of Technology. He is currently a Professor with Heilongjiang International University. He has published more than 50 academic articles at home and abroad, and has been indexed and collected more than 20 articles by SCI and EI. His research interests include data mining, intelligent transportation, big data, social computing, intelligent medical, and sports engineering. He is a member of ACM and also an Advanced Member of CCF.

YANWEI WANG received the Ph.D. degree from Harbin Engineering University. She is currently a Professor and a Visiting Scholar with the School of Mechanical Engineering, Purdue University. She has published 24 research articles in scholar journals in the above research areas and has participated in several books. Her research interests include image processing, PIV/microPIV, and gas measurement.

DONG LI received the Ph.D. degree from Shanghai Jiao Tong University. He is currently a Professor with the School of Computer Science, Harbin Institute of Technology. His research interests include computer network, information security, and computer system structure.