IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 4, 2019, accepted December 14, 2019, date of publication December 17, 2019, date of current version January 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2960456

Structured Medical Pathology Data Hiding Information Association Mining Algorithm Based on Optimized Convolutional Neural Network

XIAOFENG LI[1], (Member, IEEE), YANWEI WANG[2], AND GANG LIU[3]

[1]Department of Information Engineering, Heilongjiang International University, Harbin 150025, China
[2]Department of Mechanical Engineering, Harbin Institute of Petroleum, Harbin 150027, China
[3]College of Computer Science and Technology, Harbin Engineering University, 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 When using traditional algorithms to mine the association of hiding information in medical pathological data, there are some problems, such as low recognition rate of association and poor accuracy of mining results. Therefore, structured medical pathology data hiding information association mining algorithm based on optimized convolution neural network is proposed. Firstly, an information feature is optimized based on rough set relative classification information entropy and ant colony algorithm and the optimized feature matrix is obtained. The information in the optimized feature matrix is weighted, and the weighted features of hiding information are obtained. Secondly, the hiding information feature matrix is transmitted to the convolution neural network for learning, and the weight of the connection layer is extracted. The importance of the corresponding area of the weight is confirmed by the distribution of the weight value, and the feature average matrix is obtained. According to the matrix, the feature of hiding information data is enhanced. The hiding information in the structured medical pathology data is generalized by using the Gaussian Bell function, and the hiding information generalization processing result is combined with the adjacent matrix in the convolution neural network to construct the hiding information classification model. Finally, the classification standard is defined, the cooperative association of hiding information group is obtained, and the mining of association between hiding information of structured medical pathological data is completed. The experimental results show that the proposed algorithm has good feature optimization effect, and the information association recognition rate is high, the anti-interference ability and accuracy are better than the current related results, the highest recall rate is 99.24%, which is much higher than the traditional algorithm, which shows that the algorithm is effective.

INDEX TERMS Convolutional neural network, deep learning, medical pathology data, hiding information, association mining.

I. INTRODUCTION

As a key technology for information analysis in various fields of society, data mining is indispensable in the development of various industries [1]. In foreign countries, information data mining is widely applied in retail, biology and information retrieval [2], [3]. The Children's Hospital in Chicago, USA, has built a data mining system based on SPSS to achieve prediction and analysis, and assist in the treatment of childhood brain tumors, thereby effectively reducing the overall operating costs of the hospital and reducing the cost of patient treatment [4], [5]. In China, most of the data mining technology is applied in large enterprises, and the application is still relatively less in the medical industry [6], [7]. However, through the continuous efforts of the personnel in this field, some excellent research results have been obtained [8], [9].

In order to deal with issues such as poor large-scale data mining in the medical field, Literature [10] proposed a new data collection and mining strategy, analyzed the relevance of obstetric information data in current medical institutions based on Apriori algorithm, focused on the association between cesarean section and the existing signs and the drugs

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

used, and analyzed the association between maternal hospitalization and the time and number of births. In addition, analyzed the medical system related to data acquisition and transmission rate, used mobile portable digital terminal to achieve data acquisition, and expounded the distributed multi-layer system and SSH2 framework construction strategy. Literature [11] proposed a mining algorithm that did not generate candidate sets, used a new tree structure to store sliding window information, established a utility database corresponding to the stored data information, and calculated the utility of all information in the window, thereby effectively solving the frequent problem of data item sets. Literature [12] studied the local pattern mining problem in gene expression data, and transformed the analysis of gene expression data from the original analysis from the overall mode to the analysis from the local mode, so that the situation of clustering data only according to all objects or attributes of data was improved; in addition the literature introduced the research status and progress of local pattern mining in gene expression data, and analyzed the local pattern mining standards; the experimental results show that the method could mine the association information in the data. However, the above methods have the defects of poor data feature optimization, low recognition rate, vulnerability to interference factors, and data mining results are not accurate.

According to the above analysis, when mining the association of the hiding information of structured medical pathological data, the candidate patterns existing in a large amount of information should be fully considered, and the interference factors generated during the mining process should be controlled to ensure the accuracy of information mining and improve the application value of the algorithm.

In order to solve issues such as low recognition rate and poor accuracy of mining results in traditional methods, this paper used the feature that Convolutional Neural Network (CNN) had similar structure and group structure in the running process, and proposed the a structured medical pathological data hiding information association mining algorithm based on optimized CNN. The algorithm solved the problems of low recognition rate of hiding information and poor accuracy of mining results in traditional methods. The experimental results show that the proposed algorithm can effectively improve the recall rate of data association, and the accuracy of mining results is high, which can solve the current problems and meet the actual needs. The main contributions of this paper include four aspects:

1. An information feature optimization method based on rough set relative classification information entropy and ant colony algorithm, which was used to ensure the consistency of information features.

2. A hiding information generalization processing method based on Gaussian Bell function was proposed. After generalization processing, the accuracy of data mining could be guaranteed, thereby improving the shortcomings of data mining accuracy in traditional methods.

3. On the basis of information feature optimization and information generalization processing, the optimized CNN was applied to the structured information mining data hiding information association mining, and the feature data matrix was enhanced by the feature average matrix to improve the data hiding information association recognition rate.

4. The experimental results show that the proposed algorithm can produce larger information feature optimization evaluation factor and better results, and effectively improve the recognition rate of data association, anti-interference ability, recall rate and data association mining accuracy. Therefore, the proposed algorithm has effectiveness.

II. RELATED WORK

The intrinsic attribute analysis of structured pathological data and the mining of hiding information have a very important guiding role in medical diagnosis [13]–[15]. In medical diagnosis, the intrinsic mechanism of disease is complicated and the relationship between disease and human body is also very complex, and there is usually a certain nonlinear association [16]. If many diseases are diagnosed based on professional knowledge, it is difficult to find hiding pathological information [17], [18]. With the development of medical informatization, the use of computers to mine a large number of information hiding in pathological data has very significant results.

Literature [19] analyzed the current social network situation, and showed the necessity of related user mining research; the related user mining was analyzed from three aspects: user attribute, user relationship and its synthesis, but the proposed method was ineffective in optimization of data characteristics. Literature [20] proposed an association rule mining method, considered the dynamic data characteristics in the market, helped to identify user behavior well and completed data mining, but the data recognition rate and recall rate were low. Literature [21] proposed a sensitive association rule method based on confidence for some important privacy data; the main goal of the proposed method was to hide a set of interesting patterns, which contained sensitive knowledge and minimal knowledge of side effects. The experimental results show that the proposed method is effective in terms of rule loss, modification times, hiding faults and complete avoidance of ghost rules; but it also produced a large number of candidate patterns during the mining process, and the candidate pattern generation needs to be compared multiple times; as a result, this method had the deficiencies of low data association recognition rate and poor accuracy [22].

This paper took advantage of computer technology to analyze the hiding information of structured medical pathological data, and implemented the research based on the optimized CNN. CNN has a simple structure and can solve image analysis problems in the field of computer vision. Some scholars skillfully used the end-to-end mapping between low resolution and high resolution, proposed super-resolution CNN, and focused on analyzing the learning function of the hiding layer, thereby greatly improving the efficiency of the

algorithm [23]. However, due to the increasing and deepening of the image types, the performance of CNN convergence speed has been affected, and CNN needs further optimization.

Based on the optimized CNN, the hiding information of structured medical pathological data was correlated and mined. The information features were optimized before the pathological data was input into CNN. Compared with the existing literature methods, this method has good mining performance.

III. STRUCTURED MEDICAL PATHOLOGY DATA HIDING INFORMATION ASSOCIATION MINING ALGORITHM

A. INFORMATION FEATURE OPTIMIZATION METHOD BASED ON ROUGH SET RELATIVE CLASSIFICATION INFORMATION ENTROPY AND ANT COLONY ALGORITHM

For the multi-class data with complex medical pathological data characteristics, it is difficult to mine the association information of hiding information [24]. The rows and columns of the data are in a state of mutual connection, if several rows or columns are selected separately, the optimization features, the recognition rate, and the hiding information association mining can not be obtained.

At the same time, in the case of complex features, there will be a lot of redundant information in the data, namely, different regions in the data structure network show different importance [25]. In summary, information feature optimization method based on rough set relative classification information entropy and ant colony algorithm, the application of the proposed method to information hiding in pathological data feature optimization lays the foundation for information association mining.

To define the fitness function suitable for information feature optimization, and use $DT = (U, A \bigcup D, V)$ to represent the decision table, V is the feature evaluation matrix. Original data set $B \subseteq A$, then the probability distribution of information feature U/B relative to U/D is:

$$f_{ij} = \frac{|B \cap D|}{D} \tag{1}$$

The relative classification information entropy of U/B relative to U/D can be expressed as:

$$G(U/B) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{|B_i|}{U} f \tag{2}$$

Among them, the relative classification information entropy can reflect the classification results of information features, and express the inconsistency of information features. The degree of inconsistency reflected by information entropy is proportional to the information eigenvalue: the smaller the information entropy, the higher the information consistency. Therefore, this paper selected the fitness function, or the relative classification information entropy defined by Formula (2), so as to optimize the information features and obtain the feature matrix [26].

This Section used a binary ant colony algorithm. The solution to the problem could be represented by multiple dimensions, and the optimized feature matrix calculation was performed based on the above-mentioned feature matrix. The process is as follows:

1) Randomly generate several vectors as the initial ant colony, and use Formula (2) to calculate the classification information entropy of each ant colony;

2) Update the historical optimized value of each ant colony, compare all ant colonies, find the global optimum, and obtain the optimized speed and optimized position relationship of the ant colony. The calculation formula of optimized speed is as follows:

$$v = v^k + rand(q^k - q_0) + rand(v^k - v_0) \tag{3}$$

where, v_0 is the original velocity of the ant colony, v^k is the speed at which the $k-$th ant colony moves, q_0 is the original position of the ant colony, and q^k is the position of the $k-$th ant colony movement.

3) Assign an ant colony feature weight according to the optimized speed and the optimized position, thereby obtaining a feature matrix corresponding to each ant colony;

4) After the iteration is completed, output the optimized feature matrix.

Combined with the ant colony algorithm, the optimized feature matrix is searched by the ant colony global optimized motion step. Based on this, the information in the optimized feature matrix is weighted to obtain the weighted features of the hiding information:

$$\chi_i = \log(f_i + 1) \times \left(1 + \log \frac{1}{N} \sum_{i=1}^{N} \left[\frac{f_{ik}}{n_i}\right]\right) \tag{4}$$

where, χ is the weighted result of the information, and f is the frequency of occurrence of hiding information in the sample data. n is hiding information, N is the number of ant colonies.

On the basis of the optimization of the above information features, the feature matrix is transmitted to the CNN for learning, so as to extract the weight of the connection layer, and use the obtained weight value distribution to confirm the importance degree of the corresponding region of the weight, and obtain the feature average matrix. At the same time, the enhancement of the hiding information data features is implemented according to the matrix [27]–[29]. The detailed process is as follows:

The original information data D is set as a $d_1 \times d_2$ matrix, which is convoluted by multiple layers and input into the matrix of the $s \times n \times l$ in the CNN fully connected layer, where s is the number of input features and the feature dimension is $n \times l$. At the same time, there is an t-layer fully connected layer in the CNN, and the full-connection layer of each layer can be described by $W_1 \cdots W_t$, and the number of nodes can be represented as $k_1 \cdots k_t$. There are k_t nodes in the t-th layer fully connected layer, indicating k_t classification problems. j is the minimum number of nodes

The average value of the beneficial weighting values for all classifications is:

$$\bar{W}(j) = \frac{1}{k_t}\sum_{j=1}^{k_t} W(j) \tag{5}$$

Calculate according to Formula (5), rearrange $\bar{W}$ to obtain a matrix of W' sized $[s \times n \times l]$:

$$W' = [s \times n \times l, 1] \rightarrow [s \times n \times l] \tag{6}$$

According to the above calculation, W' can be regarded as the weight data of $sn \times l$. Here, by taking an absolute value for each data, it is possible to obtain a deviation of the mean values of all the values in one data region, thereby enabling the effective weights to be converted into large values, and the invalid weights will infinitely approach [30], [31]. Do the following operations for the t-th data:

$$W''(t) = \left| W'(t) - \frac{1}{s \times l}\sum_{t=1}^{k_t} W'(t) \right| \tag{7}$$

Calculate the average of the first dimension, then there is:

$$T(t) = \frac{1}{s}\sum_{t=1}^{k_t} W''(t) \tag{8}$$

For T, whose dimension is $n \times l$, the operation of evaluating the original data D is described. The random raw data (x, y) is the evaluation of the region $((x-1) \times (d_1/n - x \times d_1/n)$, $(y-1)d_2/l - y \times d_2/l)$ in D.

Combining the above calculations and analysis, the feature evaluation matrix V with the same D dimension can be generated by using the definition of T, wherein each element in V describes the element evaluation in the corresponding D.

For the optimization of hiding information features of the structured medical pathological data, the above process can be used to obtain a data set after feature optimization, and then apply the set to the information association mining [32], which can effectively improve the mining quality.

B. HIDING INFORMATION GENERALIZATION PROCESSING METHOD BASED ON GAUSS BELL FUNCTION

According to the optimization result of information features, the Gaussian Bell function was used to generalize the hiding information in structured medical pathological data. The generalization value is:

$$\text{Generalization value} = [1, Num\,\mathrm{Pr}\,ototypes] : \varphi(X_k) > e^{-1} \tag{9}$$

where, $\varphi(X_k)$ is the eigenvalue obtained by decomposition of the medical pathological data sample, and the membership degree between the sample information and the original information is obtained by combining the membership function, which can be calculated by Formula (10).

$$\varphi_i(X_k) = e\left[\frac{\cos Dist(\mathrm{Pr}\,ot, z_k}{\varepsilon_i}\right] \tag{10}$$

where, $i = [1, Num\,\mathrm{Pr}\,ot]$, ε_i is an extension of the membership function.

According to Formula (9), the dispersibility of the hiding information of structured medical pathological data is obtained, and the generalized result of the generalization of the scattered information is obtained as follows:

$$\varepsilon_j = \sqrt{\frac{[\varepsilon(k'-1)]}{k'}} + \sqrt{\frac{\cos Dist(k'-1)}{k'}} \tag{11}$$

where, k' is the singularity that exists in the generalization of information. The Gauss Bell function together with the membership function method are used to generalize the optimized hiding information, which can avoid the contingency and randomness in the data mining process, thus improving the reliability of information mining.

C. HIDING INFORMATION CLASSIFICATION MODEL

Based on the above hiding information optimization and information generalization results, this study used the optimized CNN to mine the association between the hiding information of the structured medical pathological data [33]. It was necessary to define a CNN first, using CNN to construct the model of hiding information classification, as detailed below.

Suppose the CNN consists of two sets of V_e and E_d, which are denoted as $G = (V_e, E_d)$. V_e is a non-empty finite set of neural network nodes, and E_d is a finite set of network edges. If the edge has a direction, it can be called a directed CNN; instead, it can be called a non-directed CNN [34], [35].

By using the concept that CNN has the similarity between structure and group structure [36], this study obtained the association between information through the adjacency matrix without obtaining the distance between medical hiding information in pathological data in advance [37], [38]. This could effectively reduce the difficulty of mining and facilitate obtaining a data adjacency matrix:

$$A_d = \begin{bmatrix} 0 & a(i,j) & \cdots & a(1,n) \\ a(2,1) & 0 & \cdots & a(2,n) \\ \cdots & \cdots & \ddots & \cdots \\ a(n,1) & a(n,2) & \cdots & 0 \end{bmatrix} \tag{12}$$

where, if the i-th data is the parent of the j-th data, then $a(i,j)$ is 1; otherwise, $a(i,j)$ is 0. Assuming that the data has a single parent node, the data classification model or the association mining model is:

$$x_{k+1,i} = A_d F_{k,h} x_{k,h} + b_k(h,i) + B_{k,i} w_{k,i} \tag{13}$$

$$z_{k+1,i} = C_{k+1} x_{k+1,i} + v_{k+1,i} \tag{14}$$

where, $x_{k,i} \in X_k$ is the set of position and velocity of data i on the horizontal and vertical axes, and i is the parent of data i. $b_k(h,i)$ is a compensation vector, which is the positional relationship between the data Q and its parent node; F and C are the data vector transfer matrix and the observation matrix,

respectively; B is the data vector noise matrix; w and v represent the noise present in the CNN, respectively. The observed noise is subject to the positive distribution [39]–[41].

By analyzing the adjacency matrix, it is easy to confirm the association between the information in the medical information hiding in the pathological data group, such as the parent-child association [42]–[44]. Assuming there is no parent node, the data is called a head node. The classification of such nodes will have an impact on the data, and the classification of the head node itself will not be affected by other data [45]–[48]. Therefore, the compensation vector b in the head node classification is defined as 0, and $x_{k,h}$ is its own vector at time k; otherwise, the data has a parent node, and the data classification is affected by the parent node, so the data classification model is utilized [49]. It is possible to identify that the compensation vector b contains detailed information about the direction and distance between the node and its parent node. If the information data has multiple parent nodes, then under the linear conditions, $x_{k+1,i}$ can be defined as:

$$x_{k+1,i} = \sum w_k\left(h,i\right)\left[F_{k,h}x_{k,h} + b_k\left(h,i\right)\right] + B_{k,i}w_{k,i} \quad (15)$$

where, $x_{k,i} \in X_k$, $\sum w_k\left(h,i\right) = 1$, $w_k\left(h,i\right) \in [0,1)$.

Based on the above CNN, the information data adjacency matrix can be obtained, Data1 is the head node, and Data2 and Data3 are the child nodes of Data1, then the information data group classification model [50], [51], that is, the association model can be expressed as:

$$\begin{cases} x_{k+1,1} = F_{k,1}x_{k,1} + B_{k,1}w_{k,1} \\ x_{k+1,2} = F_{k,2}x_{k,1} + b_k\left(1,2\right) + B_{k,2}w_{k,2} \\ x_{k+1,3} = F_{k,3}x_{k,1} + b_k\left(1,3\right) + B_{k,3}w_{k,3} \end{cases} \quad (16)$$

D. HIDING INFORMATION ASSOCIATION MINING ALGORITHM FOR STRUCTURED MEDICAL PATHOLOGICAL DATA

Based on the information data group classification model obtained from the above section, a classification standard is defined, and the hiding information association mining of structured medical pathology data is realized by CNN.

The vectors between the information are correlated and not independent. However, since the coordination association between the initial stages of information cannot be obtained, the relationship between the information can be defined as the coupling relationship between the group structure and the information word vector. Here, the optimized CNN algorithm can be divided into two stages. In information mining, information should be treated as an independent classification.The specific mining algorithm can be described as follows:

Input: test samples and training samples based on the hiding information to be mined, by formula (16) to obtain information data group classification model $x_{k,i}$

Output: Result of mining hiding information of structured medical pathology data.

Initialize structured medical pathological data, and perform pathological data hiding information association mining based on optimized convolutional neural network.

Setp1: In this study, a classification standard was defined when information was mined by CNN:

$$\pi\left(X\right) = \Delta\left(X\right)\sum_{c'}\omega^{c'}\left(I\right)\left(F\left(L\right)\right)p^{c'} \quad (17)$$

where, c' is a discrete variable, $\omega^{c'}\left(I\right)$ is an information weight value, $F\left(L\right)$ is a set of all finite information subsets, and $p^{c'}$ is an information probability density.

Setp2: Formula (17) is in a closed state under Bayesian recursion. When the prior probability density of the hiding information of the structured medical pathological data is in the form of Formula (17), then the simplified form prediction process of the information classification standard.

$$\pi\left(X'\right) = \Delta\left(X'\right)\sum_{c'}\omega^{(I,\zeta)}p^{\zeta} \quad (18)$$

$$\omega^{(I,\zeta)} = \omega_B\left(I\cap B\right)\omega_s^{\xi}\left(I\cap L\right) \quad (19)$$

$$p^{\zeta} = = p_s^{\zeta}\left(x,\lambda\right) + p_B\left(x,\lambda\right) \quad (20)$$

where, $\omega_B\left(I\cap B\right)$ is the weight value of the newly added information label, $\omega_s^{\xi}\left(I\cap L\right)$ is the weight value of the existing label, $p_s^{\zeta}\left(x,\lambda\right)$ is the density of the maintenance information acquired by the prior density, and $p_B\left(x,\lambda\right)$ is the probability density of the newly added information.

Setp3: Assuming that the predicted density of multiple pieces of information is not the previous prediction result, the classification criteria can be expressed in the form of Formula (21), and Formula (21) is the information-associated classification model under the current conditions, namely the mining model:

$$\pi\left(X\,|z\right) = \delta\Delta\left(X'\right)\sum_{(I,\zeta)\in F(L)\times G}\sum_{\Theta^M}\omega^{(I,\zeta,\theta)}D'p^{\zeta} \quad (21)$$

where, Θ^M is a set of M elements at the maximum weight, and $\omega^{(I,\zeta,\theta)}$ is the information association normalization weight value, which is determined based on the expert method, according to the expert's knowledge and experience to determine the weights of the information indicators in all directions, calculate the mean value and standard deviation of each indicator weight, and determine the final weight by continuous feedback and modification.

Setp4: According to the obtained information word vector mining results, the classification group structure is obtained, and then the information group cooperation association is obtained, thereby completing the mining of the association information between the structured medical pathological data hiding information.

Setp5: End

Based on the above analysis, the hiding information association mining of structured medical pathological data is realized. The overall mining process can be shown in Figure1.

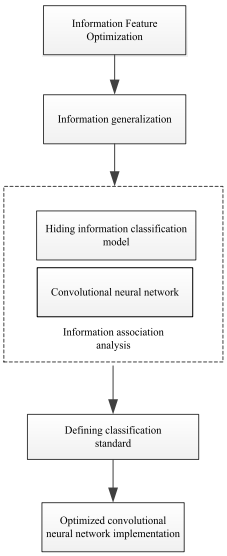

FIGURE 1. Overall mining process flow.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the overall effectiveness of the hiding information association algorithm based on optimized CNN structured medical pathological data, a comparative experiment was conducted in this study.

A. EXPERIMENTAL DATA

The experimental data sources from TCGA (The Cancer Genome Atlas)(https://portal.gdc.cancer.gov/).The TCGA database contains clinical data, genomic variation, mRNA expression, miRNA expression, methylation and other data of various human cancers. It is an important data source for cancer researchers.Ten million pathological data sets were extracted from TCGA database. The experimental data set mainly includes categories such as BreaKHis data set, Camelyon data set, Original sequencing data set, and Differential gene data set. Two point five million data collected per category. The information hiding in pathological data association was mined using the optimized CNN algorithm. In the experiment, 70% of the data was used as training data, 30% of the data was used as test samples, and the experimental platform was Matlab.

B. EXPERIMENTAL INDICATORS

The experimental data sets were all standard formats and could be directly applied to experiments. Based on the above experimental environment and data, five experimental indicators were selected for experimental analysis, as follows:

1) RESULTS OF FEATURE OPTIMIZATION EVALUATION

As Section III(A) pointed out, in order to avoid the problem of information association mining due to information redundancy under complex conditions, this paper completed the optimization of information features based on rough classification information entropy and ant colony algorithm. In order to verify the feature optimization effect, this study evaluated the feature optimization results, according to formula (8), it can be known that the average value T of the first dimension describes the evaluation of the original data and uses it as a feature optimization factor to reflect the advantages and disadvantages of the evaluation results. The larger the value T, the better the result of feature optimization;

2) ASSOCIATION RECOGNITION RATE OF HIDING INFORMATION

In the implementation process of the proposed algorithm, the association between nodes needed to be recognized, and the proposed algorithm was compared with the methods in Literature [5], Literature [6], Literature [8] and Literature [9]. The calculation formula of the hiding information association recognition rate is as follows:

$$Recog = \frac{J}{All\ right\ information} \times 100\% \tag{22}$$

where, *All right information* is the true amount of hiding information. J is the amount of hiding information identified by the algorithm.

3) ANTI-INTERFERENCE PERFORMANCE OF DATA CLASSIFICATION RESULTS

The construction of information group classification model is an important condition for realizing data mining. As pointed out in Section III(C), the data classification results may be affected by the nodes, so this study analyzed the anti-interference performance of the data classification results and expressed with anti-interference factor τ in formula(16); the larger the τ, the stronger the anti-interference ability of the algorithm is. The value range is set to [0-1];

4) MINING RECALL RATE OF DATA ASSOCIATION

The recall rate of the association mining algorithm was compared to verify the performance of the algorithm. The formula for calculating the recall rate is as follows:

$$Rec = \frac{Information\ extracted}{All\ right\ information} \times 100\% \tag{23}$$

where, *information extracted* is the amount of hiding information extracted.

(5) Accuracy of Hiding Information Association Mining
The proposed algorithm was compared with the methods in Literature [6], Literature [7], Literature [8] and Literature [9] and Literature [10] in terms of mining accuracy, thereby verifying the advantages of the proposed algorithm.

$$Acc = \frac{Correct\ information}{All\ right\ information} \times 100\% \tag{24}$$

where, *Correct information* is the correct amount of hiding information

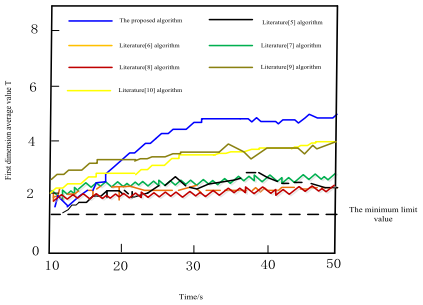

FIGURE 2. Comparison of feature optimization evaluation results of different methods.

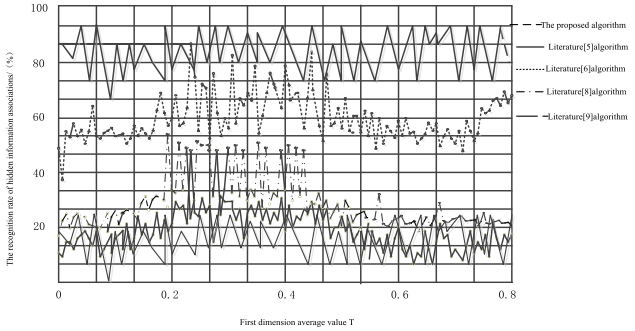

FIGURE 3. The influence of the first dimensional average value on the recognition rate of hiding information association under different methods.

C. COMPARISON OF FEATURE OPTIMIZATION EVALUATION RESULTS

The comparison results of the optimization evaluation of the information features under different methods, it is shown in Figure2.

According to the analysis result of Figure2, the first dimensional average value of the above methods were all higher than the minimum limit value. The first dimensional average value of Literature [5], Literature [6], and Literal [7] and Literature [8] were around 2, and the first dimensional average value of the Literature [9] and Literature [10] methods were relatively high up to 4 but still below the proposed algorithm. In comparison, the first dimensional average value of the proposed algorithm was significantly higher than the traditional method, and kept increasing with the increase of experimental time. This is because the proposed algorithm is used to optimize the feature of the hiding information based on the rough classification relative index information entropy and ant colony algorithm, which improves the feature optimization effect and provides the basis for the hiding information association mining.

D. COMPARISON OF HIDING INFORMATION ASSOCIATION RECOGNITION RATE

The comparison results of the hiding information association recognition rates of different methods are shown in Figure3.

According to Figure3, as the value of the first dimensional average value T changes, the association recognition rate of the hiding information also changes, and the overall trend is fluctuating. Compared with the literature results, the association mining algorithm based on the hiding information of the structured medical pathological data optimized by CNN shows a high recognition rate of data association, which fluctuates around 80%. The method of Literature [6] can reach 80% at the peak, but the average level is about 60%. The methods of the Literature [5], Literature [8] and Literature [9] ranged from 8% to 40%. It can be seen that the first dimensional average value will have an impact on the mining effect. In the data mining, the influence of the feature shadow should be fully considered. The proposed algorithm takes into account the role of the first dimensional average value T.

TABLE 1. Comparison of anti-interference performance of data classification results.

Algorithm	Influence of nodes τ value	influence of no nodes τ value
Literature[6] algorithm	0.52	0.51
Literature[7] algorithm	0.25	0.62
Literature[8] algorithm	0.61	0.84
Literature[9] algorithm	0.52	0.77
Literature[10] algorithm	0.20	0.45
The proposed algorithm	0.94	0.98

Before mining the medical information hiding in pathological data association, the information features were optimized, and substitutes it into the CNN data classification model, thereby realizing the hiding information association mining, improving the information recognition rate and enhancing the quality of data mining.

E. ANTI-INTERFERENCE PERFORMANCE OF DATA CLASSIFICATION RESULTS

Comparison of anti- interference performance between literatures results and the data classification results of proposed algorithm under the influence of node and no node. It is shown in Table1.

According to the analysis data of Table 1, the above several methods are affected by the nodes. Under the influence of no nodes, the τ value is large and the anti-interference ability is strong. The method of Literature [10] has the smallest τ value under the influence of nodes, which is 0.20, followed by Literature [7], τ value is 0.25, and the method of Literature [8] and Literature [9] has strong resistance to nodes of 0.61 and 0.52 respectively, but still low. In the proposed algorithm and under the condition of node influence, the τ value of the proposed algorithm is above 0.90, which indicates that the data classification model constructed in this paper has better performance and provides a good foundation for data association mining. This is mainly due to the fact that this paper first generalizes the information, and avoids the contingency and randomness in the data processing process, thereby realizing data classification, reducing external interference and enhancing classification performance.

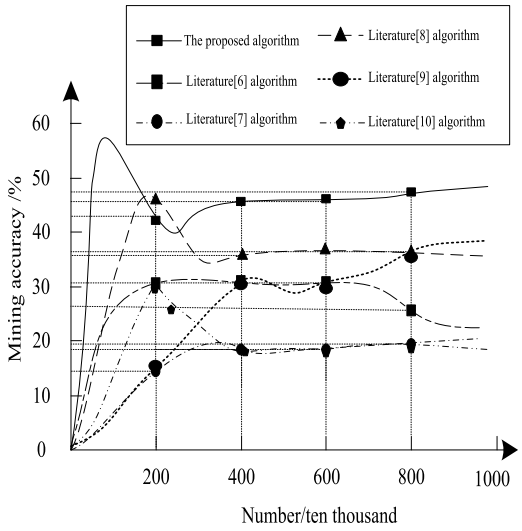

FIGURE 4. Comparison results of data association mining accuracy under different methods.

F. COMPARISON OF THE ACCURACY OF HIDING INFORMATION ASSOCIATION MINING

In order to further verify the experimental results of Section IV (E) and the effectiveness of the proposed algorithm, this study compared the accuracy of hiding information association mining of structured medical pathological data between the proposed algorithm and the methods in Literature [6], Literature [7], Literature [8], Literature [9], and Literature [10]. The comparison results are shown in Figure4.

According to Figure4, the mining accuracy of the proposed algorithm is always at a high level. When the data amount is one million, the highest mining accuracy reaches 79%. After the data amount is two million, the accuracy slightly decreases, which is lower than Literature [8], but it quickly rises going up, after the data amount is two point six million, the mining accuracy of the proposed algorithm is always higher than that of several other literatures.

It can be seen from Figure4 that the accuracy of Literature [9] shows an upward trend, up to 40%. The accuracy of Literature [7] and Literature [10] rises first and then decreases, at about 20%. The accuracy of Literature [6] is up to 53%. The mining accuracy of Literature [8] is up to 60%, which is the highest accuracy among several literatures. However, compared with the proposed algorithm, there is still a large gap.

According to the above data,when the proposed algorithm is used to mine the hiding information association in structured medical pathological data, the mining accuracy is higher than the traditional method, which shows that the results obtained by this algorithm are more accurate and can be applied in practice.

G. COMPARISON OF RECALL RATES OF DATA ASSOCIATION MINING

Table2 compared the mean of the information and the recall rate of the information data association data of the proposed algorithm under different experimental times.

TABLE 2. Data relevance mining recall rate comparison of different research results.

Number of experiments	Literature [7]/%	Literature [8]/%	Literature [9]/%	Literature [10]/%	The proposed algoritm/ %
10	85.23	90.13	83.23	90.32	97.68
20	87.24	82.27	82.35	89.32	99.24
30	86.59	80.28	82.25	93.50	97.58
40	87.23	90.02	71.02	91.02	98.15
50	86.15	92.24	73.02	92.21	98.45
60	85.32	90.72	79.36	93.56	97.29
70	84.98	80.38	78.25	87.02	98.26
80	85.03	89.51	80.21	89.36	96.06
90	84.36	82.10	81.36	90.21	95.37
100	85.20	90.04	86.35	94.02	97.35

Data analysis results in Table2 shows that the highest recall rate of the structured information path mining data based on the optimized CNN is 99.24%, and the highest recall rate of Literature [7] is 87.24%, the recall rate of Literature [8] is up to 92.24%, the highest recall rate of Literature [9] is 86.35%, and the highest of Literature [10] is 94.02%.

By comparison, the proposed algorithm's data association mining recall rate is significantly higher than that of Literature [7] and Literature [8], indicating that the proposed algorithm has superiority in effectively mining the association information of structured medical pathological data and the data mining quality of the algorithm is high.

In summary, the experimental results show that the proposed method has better feature optimization and it is better than the traditional methods in terms of data association recognition rate, data association mining recall rate and data association mining accuracy, indicating that the proposed algorithm can realize the effective mining of hiding information association of structured medical pathological data after the CNN is optimized.

V. CONCLUSION

In order to solve the problem of low recognition rate and poor accuracy of mining results of traditional algorithm, this study used the optimized CNN to analyze the hiding information association mining of structured medical pathological data. First of all, this study optimized the information features to highlight the significant features of the information and reduce the data redundancy rate. Then, the CNN was used to realize the information mining classification mining. The experimental results of the proposed algorithm show that the proposed algorithm is superior in performance and feasible. It can provide some reference for data mining research. Although the proposed algorithm has achieved some results, there are still many shortcomings, and the application of convolutional neural networks is insufficient. In the future, we should fully study the convolutional neural network and use its powerful computing advantages to further analyze the data mining hierarchy to get more accurate mining results.

REFERENCES

[1] T. Linbo, S. Jianjing, and Y. Qingxiang, "Cloud data privacy protection in data mining," *Comput. Sci.*, vol. 43, no. 5, pp. 113–116, 2016.

[2] W. Yishu, Y. Ye, and L. Meng, "Overview of query processing and mining technology for large-scale time series map data," *Comput. Res. Develop.*, vol. 55, no. 9, pp. 65–78, 2018.
[3] W. Ruojia, W. Siyi, and Z. Yiran, "Review of the application of data mining in the field of health care," *Library Inf. Knowl.*, vol. 185, no. 5, pp. 116–125, 2018.
[4] H. Gao, W. Huang, and X. Yang, "Applying probabilistic model checking to path planning in an intelligent transportation system using mobility trajectories and their statistical data," *Intell. Automat. Soft Comput.*, vol. 25, no. 3, pp. 547–559, Jan. 2019.
[5] N. T. Rao, "A review on industrial applications of machine learning," *Int. J. Disaster Recovery Bus. Continuity*, vol. 8, pp. 1–10, Nov. 2018.
[6] E. Rashid, "Disease detection on the basis of multiple symptoms by expert system," *Int. J. Disaster Recovery Bus. Continuity*, vol. 8, pp. 1–10, Nov. 2017.
[7] 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.
[8] Y. Yin, W. Xu, Y. Xu, H. Li, and L. Yu, "Collaborative QoS prediction for mobile service with data filtering and SlopeOne model," *Mobile Inf. Syst.*, vol. 2017, pp. 7356213:1–7356213:14, Jun. 2017.
[9] 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.
[10] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.
[11] G. G. Shiming, "An efficient algorithm for mining efficient item sets of data streams based on sliding windows," *J. Harbin Univ. Eng.*, vol. 39, no. 4, pp. 721–729, 2018.
[12] J. Tao and L. Zhanhuai, "A review of local pattern mining in gene expression data," *Comput. Res. Develop.*, vol. 55, no. 11, pp. 3–20, 2018.
[13] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
[14] G. Jia, G. Han, H. Rao, and L. Shu, "Edge computing-based intelligent manhole cover management system for smart cities," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1648–1656, Jun. 2018.
[15] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 767–779, Apr. 2015.
[16] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 5, pp. 1005–1016, May 2017.
[17] A. Pansotra and S. P. Singh, "Additive Hough transform and fuzzy C-means based lane detection system," *Int. J. Disaster Recovery Bus. Continuity*, vol. 8, pp. 11–28, Nov. 2017.
[18] K. S. Rao, K. V. Satyanarayana, and P. S. Rao, "Segmentation of images using two parameter logistic type distribution and K-means clustering," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 20, pp. 1–20, 2018.
[19] Z. Xiaoping, L. Xun, and Z. Jichao, "Overview of methods of mining associated users for social network convergence," *J. Softw.*, vol. 28, no. 6, pp. 1565–1583, 2017.
[20] M. Kaur and S. Kang, "Market basket analysis: Identify the changing trends of market data using association rule mining," *Procedia Comput. Sci.*, vol. 85, pp. 78–85, Jan. 2016.
[21] S. Krishnamoorthy, G. S. Sadasivam, M. Rajalakshmi, K. Kowsalyaa, and M. Dhivya, "Privacy preserving fuzzy association rule mining in data clusters using particle swarm optimization," *Int. J. Intell. Inf. Technol.*, vol. 13, no. 2, pp. 1–20, 2017.
[22] M. Vardhana, "Fuzzy logic based performance analysis of various antenna structures," *Int. J. Grid Distrib. Comput.*, vol. 12, no. 1, pp. 1–10, 2019.
[23] Y. Yin, S. Aihua, G. Min, X. Yueshen, and W. Shuoping, "QoS prediction for Web service recommendation with network location-aware neighbor selection," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 26, no. 4, pp. 611–632, 2016.
[24] G. Jia, G. Han, J. Jiang, S. Chan, and Y. Liu, "Dynamic cloud resource management for efficient media applications in mobile computing environments," *Pers. Ubiquitous Comput.*, vol. 22, no. 3, pp. 561–573, 2018.
[25] H. Si, Y. Qi, M. Zheng, Y. Ren, and L. Yu, "Structured peer-to-peer-based publication and sharing of ontologies to automatically process SPARQL query on a semantic sensor network," *Int. J. Distrib. Sensor Netw.*, vol. 14, no. 10, 2018, Art. no. 155014771879758.
[26] X. Li, Y. Wang, and D. Li, "Medical data stream distribution pattern association rule mining algorithm based on density estimation," *IEEE Access*, vol. 7, no. 1, pp. 141319–141329, 2019.
[27] A. Anguera, J. M. Barreiro, and J. A. Lara, "Applying data mining techniques to medical time series: An empirical case study in electroencephalography and stabilometry," *Comput. Struct. Biotechnol. J.*, vol. 14, pp. 185–199, Jan. 2016.
[28] M. Takada, M. Fujimoto, and K. Hosomi, "Association between benzodiazepine use and dementia: Data mining of different medical databases," *Int. J. Med. Sci.*, vol. 13, no. 11, pp. 825–834, 2016.
[29] G. Tzanis, "Biological and medical big data mining," *Int. J. Knowl. Discovery Bioinf.*, vol. 4, no. 1, pp. 42–56, 2017.
[30] T. Si and S. Sujauddin, "A comparison of grammatical bee colony and neural networks in medical data mining," *Int. J. Comput. Appl.*, vol. 134, no. 6, pp. 1–4, 2016.
[31] R. J. Oskouei, N. M. Kor, and S. A. Maleki, "Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges," *Amer. J. Cancer Res.*, vol. 7, no. 3, pp. 610–627, 2017.
[32] 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.
[33] J. C. Drees, H. Karl, and M. S. Petrie, "Reference intervals generated by electronic medical record data mining with clinical exclusions: Age-specific intervals for thyroid-stimulating hormone from 33038 euthyroid patients," *J. Appl. Lab. Med.*, vol. 3, no. 2, pp. 231–239, 2018.
[34] K. K. L. Wong, D. Wang, and S. Fong, "A special section on advanced computing techniques for machine learning and data mining in medical informatics," *J. Med. Imag. Health Inform.*, vol. 6, no. 4, pp. 1052–1055, 2016.
[35] Z. Tao, R. Haijun, and H. Weijun, "A face representation learning method based on forward unsupervised convolutional neural network," *Comput. Sci.*, vol. 43, no. 6, pp. 303–307, 2016.
[36] J. Atkinson-Abutridy, C. Mellish, and S. Aitken, "Combining information extraction with genetic algorithms for text mining," *IEEE Intell. Syst.*, vol. 19, no. 3, pp. 22–30, May 2004.
[37] A. Speck-Planche and C. Mnds, "Advanced in silico approaches for drug discovery: Mining information from multiple biological and chemical data through MTK-QSBER and pt-QSPR strategies," *Current Med. Chem.*, vol. 24, no. 16, pp. 1687–1704, 2017.
[38] C. G. Daughton, "Monitoring wastewater for assessing community health: Sewage chemical-information mining (SCIM)," *Sci. Total Environ.*, vol. 619, pp. 748–764, Apr. 2017.
[39] H. Gao, D. Chu, and Y. Duan, "The probabilistic model checking based service selection method for business process modeling," *J. Softw. Eng. Knowl. Eng.*, vol. 27, no. 6, pp. 897–923, Aug. 2017.
[40] C. Yang and G. Gidófalvi, "Mining and visual exploration of closed contiguous sequential patterns in trajectories," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 7, pp. 1283–1304, 2017.
[41] C. Loglisci, "Using interactions and dynamics for mining groups of moving objects from trajectory data," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 7, pp. 1436–1468, 2017.
[42] K. Yan, Z. Minqing, and S. Tingting, "R-LWE-based multi-bit reversible information hiding algorithm in ciphertext domain," *Comput. Res. Develop.*, vol. 53, no. 10, pp. 2307–2322, 2016.
[43] L. Songbin *et al.*, "A HEVC information hiding method based on motion vector space coding," *J. Comput. Sci.*, vol. 39, no. 7, pp. 1450–1463, 2016.
[44] M. Guojun, H. Dianjun, and X. Songyan, "Large data classification model and algorithm based on distributed data stream," *J. Comput. Sci.*, vol. 40, no. 1, pp. 163–177, 2017.
[45] G. Jichang, G. Hao, and G. Chunle, "Single image deraining method based on multi-scale convolution neural network," *J. Harbin Univ. Technol.*, vol. 50, no. 3, pp. 191–197, 2018.
[46] Y. Shu, C. Fan, and H. Hongjie, "Reversible information hiding in neighborhood predictive encryption domain under XOR-scrambling framework," *Comput. Res. Develop.*, vol. 55, no. 6, pp. 97–107, 2018.
[47] N. Xinzheng, W. Chongyi, and Y. Zhijia, "An efficient association rule hiding algorithm based on cluster and threshold interval," *Comput. Res. Develop.*, vol. 54, no. 12, pp. 2785–2796, 2017.
[48] C. T. Huang, C. H. Yang, and W. J. Wang, "Raw reversibility of information hiding on the basis of VQ systems," *J. Supercomput.*, vol. 74, no. 2, pp. 1–30, 2017.

[49] H. Tian, J. Sun, C.-C. Chang, J. Qin, and Y. Chen, "Hiding information into voice-over-IP streams using adaptive bitrate modulation," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 749–752, Apr. 2017.
[50] S. A. Parah, F. Ahad, J. A. Sheikh, and G. M. Bhat, "Hiding clinical information in medical images: A new high capacity and reversible data hiding technique," *J. Biomed. Inf.*, vol. 66, pp. 214–230, Feb. 2017.
[51] H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sens. Lett.*, vol. 8, no. 5, pp. 438–447, 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, artificial intelligence, intelligent medical, and sports engineering. He is a member of ACM and 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. Her research interests include mainly in the area of image processing, PIV/micro PIV, and gas measurement. She has published 24 research articles in scholar journals in the above research areas and has participated in several books.

GANG LIU was born in 1976. He received the Ph.D. degree in computer applied technology from Harbin Engineering University, China, in 2008. He conducted research at the University of Illinois at Urbana–Champaign as a Visiting Scholar with the group of Prof. J. Han, in 2005. As a member of the China Computer Federation, he has conducted and has been conducting about ten research projects, such as the National Science and Technology Support Plan and the Chinese NSFC Project, as a Main Researcher. He is currently an Associate Professor. He has published 30 articles in well-known journals, such as JCIS, which has been cited 20 times by SCI and EI. He has authored or coauthored four books in Chinese. He has filed ten computer software copy authorities and all of them have been authorized. He has developed and applied the advanced intelligent analysis and policy consistency verification technology, in auditing more than 20 million attendees of Chinese social security.