Mobile Networks and Applications
https://doi.org/10.1007/s11036-022-02024-3

*SM 320-EMERGING CHALLENGES OF IOT AND BIG DATA TECHNOLOGIES FOR HEALTHCARE

A Deep Learning based Feature Entity Relationship Extraction Method for Telemedicine Sensing Big Data

Wenkui Zheng[1] · **Wei Hou**[1] · **Jerry Chun-Wei Lin**[2]

Accepted: 23 May 2022

Abstract
To solve the problem of inaccurate entity extraction caused by low application efficiency and big data noise in telemedicine sensing data, a deep learning-based method for entity relationship extraction in telemedicine big data is proposed. By analyzing the distribution structure of the medical sensing big data, the fuzzy function of the distribution shape is calculated and the seed relationship set is transformed by the inverse Shearlet transform. Combined with the deep learning technology, the GMM-GAN data enhancement model is built, the interactive medical sensing big data features are obtained, the association rules are matched one by one, the noiseless medical sensing data are extracted in time sequence, the feature items with the highest similarity are obtained and used as the constraint to complete the feature entity relationship extraction of the medical sensing data. The experimental results show that the extracted similarity of entity relations is more than 70%, which can handle overly long and complex sentences in telemedicine information text; the extraction time is the shortest and the volatility is low.

Keywords Deep learning · Telemedicine · Sensing big data · Features · Entity relationship · Extract

1 Introduction

Entity relation extraction is an important sub task in information extraction. Unsupervised entity relation extraction method models the task as a clustering problem, which can extract entity relations only using the information of the corpus itself. However, the existing methods are limited by the high-dimensional sparsity of entity to text co-occurrence matrix, the performance is limited and the model is complex. Although the text information is introduced into the model, the discrete feature vectors generated by the artificial feature set are also high-dimensional and sparse, which further increases the complexity of the model and reduces the improvement effect of the introduced information on the model. The rapid development of telemedicine from research to practical application has been effectively promoted through the progress of information, multimedia and communication technology, especially network communication technology. This advanced medical technology is widely used in the diagnosis and treatment of various medical specialties such as brain, chest, eye, heart, radiation and skin, as well as expert consultation of difficult and severe cases. Due to multi-agent interaction, the amount of telemedicine large data is growing, and more and more data are difficult to use effectively. How to find the relationship between entities from these unstructured texts has become an urgent need. Therefore, people have introduced deep learning algorithm to optimize telemedicine big data application system. The application field of deep learning is very wide.

For example, According to the method of reference [1], a big data abnormal risk monitoring system based on Hadoop is proposed, which shunts the big data flow, monitors the abnormal data risk by using the preprocessing end and storage end, and calculates the risk trend by using the least squares support vector machine; Fisher function is introduced to construct the uncorrelation test model; The fuzzy genetic method is used to calculate the fuzzy clustering

✉ Jerry Chun-Wei Lin
jerrylin@ieee.org

[1] College of Computer and Information Engineering, Henan University, Kaifeng, China

[2] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway

Published online: 23 August 2022

center where the abnormal data flow converges in the multi-layer space, obtain the classification gain equation of the abnormal data attribute set, and complete the big data anomaly extraction According to the experimental results, the maximum time-consuming of the proposed method is 9 s, which is significantly lower than the traditional method, and the extraction results of the proposed method are consistent with the actual situation It can be concluded that the proposed method has the advantages of high speed and high precision, and provides a technical basis for the safe transmission and application of big data However, the data noise of this method is large. Literature [2] method proposes an innovative telemedicine big data visualization method. With the explosive growth of big data, visualizing statistical data has become a challenging topic. In the past few years, it involves a lot of research work. Interpreting big data and efficiently displaying information for good understanding is a difficult task, especially in healthcare scenarios where different types of data must be managed and cross correlated. Some models and techniques of health data visualization are introduced in the literature. However, they cannot meet the visualization needs of doctors and medical staff. In this paper, we propose a new graphical tool for visualizing health data, which can be easily used to remotely monitor the health status of patients. The tool is very user-friendly and allows doctors to quickly understand the current state of patients by viewing colored circles. From a technical point of view, the proposed solution uses the geojson standard to divide the data into different circles. A recursive feature learning method for nonlinear dynamic process based on deep neural network is proposed in reference [3]. The data collected from modern industrial processes always have nonlinear and dynamic characteristics. The recently developed deep neural network method, superposition denoising automatic encoder (sdae), can extract robust nonlinear potential variables against noise from data. However, it does not consider dynamic relationships. In order to solve this problem, a new algorithm called recursive superposition denoising automatic encoder (rsdae) is proposed. In order to understand the dynamic relationship, rsdae focuses on the predictability of potential variables in the return period to include the most dynamic changes. After rsdae extracts dynamic changes, there is almost no autocorrelation in the residual. Then, the residual is monitored by principal component analysis (PCA). In order to realize process monitoring, the corresponding fault detection statistics are developed based on rsdae. Finally, a numerical example and Tennessee Eastman process benchmark verify the effectiveness of the algorithm.

Aiming at the problems of low application efficiency and large data noise of the above methods, this paper proposes a feature entity relationship extraction method of telemedicine sensing big data based on deep learning. According to the principle of gmm-gan, a data improvement model is developed. The model captures the medical sensing big data with interactive relationship, matches the association rules one by one, extracts the noiseless data, determines the features with the highest similarity, and takes this as a constraint to complete the feature extraction of telemedicine sensing big data. Based on the application efficiency and data noise of telemedicine big data, this method has achieved good performance compared with previous studies.

2 Literature review

Telemedicine sensing data is medical data that is continuously collected by wireless sensors. This data is collected, measured, and transmitted by sensing devices. Sensing devices collect a large amount of medical sensing big data in real time and dynamically in memory for backup. There is a lot of prior knowledge about the relationship between entities in telemedicine sensing big data. It is an important issue to see whether there is a relationship between entities in the knowledge base, which can provide important guidance for entity relationship extraction. To ensure that the relationship representation learned from the knowledge base encompasses the deep semantic relationship between entities, we investigate the effects of the fusion of relationship representation and text features on the performance of entity relationship extraction of telemedicine sensing big data features and analyze the telemedicine sensing big data. The sources of medical sensing big data mainly include:

1. **Complete the collection of clinical, genetic, and health data:** Develop the disease registration system, investigate the integration technology of various transmission and interface standards such as the DICOM standard and the Wado standard, realize the docking of data between various terminals, information systems, and the medical sensing big data service platform, complete the collection of medical sensing data, and provide data support for subsequent applications [4].
2. **Data transmission and preprocessing:** Transfer the data collected from the data source layer through the telemedicine network to ensure the integrity of medical sensing data as much as possible, establish the transfer and storage standards for medical sensing data, clean and preprocess the collected medical sensing data, and verify the storage format to improve the quality of data collection [5].
3. **Medical sensing big data integration provides data services for upper layer applications:** The medical sensing data is linked by managing the patient master index, and the database image, medical relational database, and medical knowledge map are set up to solve the problem of centrally storing various types of medical

sensing data. Based on Spark, MapReduce and other big data technologies, data services such as service interfaces and resource directories for upper medical applications and clinical scientific research services will be provided [6].

4. **Perform medical service applications and big data analysis services for medical research:** Develop transcriptome data analysis system, immunoprecipitation sequencing data analysis system and other precision medical applications, deploy role-based access control technology, open various kinds of business interfaces for medical institutions in the medical consortium, and finally realize the promotion and application of medical sensing data [7].
5. **Configure various types of telemedicine sensing and medical equipment:** It carries out hierarchical business cooperation, builds a medical service system with a connection between up and down, expands the amount of medical data, relies on the telemedicine network for data transmission and integrate network transmission technologies such as TB data transmission, handles fault tolerance and security, realizes the chain upgrade of telemedicine network and improves the stability and security of medical sensing data [8].

2.1 Telemedicine sensing big data extraction

The big data characteristics of the medical sensors in the database are divided into four computational groups, denoted H. The central distribution of the data features contained in the four data groups is defined as $H=(B, R, E, V)$. F represents the dimension of the global distribution of the medical data features in the database [9].With the progress of science and technology and the continuous development of technology, the amount of calculation and complexity of the algorithm are also doubling. The algorithm often contains various nonlinear operations, such as logarithmic operation, square operation, exponential operation, trigonometric function operation and so on. For example, neural network algorithm has a large number of exponential and logarithmic operations. "Logarithms and exponents are even everywhere.". Linear operation is addition and quantity multiplication. In the field of real numbers, for example, binary linear equations containing only addition and quantity multiplication belong to linear operation, such as $y=3x+5$. If it is the addition and multiplication of matrix, it is called the linear operation of matrix; If it is the addition and multiplication of vectors, it is collectively referred to as the linear operation of vectors. For different linear operations, there are generally different forms, which meet the exchange law, combination law, distribution law and so on. By performing several nonlinear calculations with the medical data features in the database, a set of high-dimensional feature spaces in this space can be obtained. Through the fuzzy association between the data features, the adaptive feature cluster center is found, and the distribution of similar features is reconstructed around the data features in the feature cluster center database to obtain the distribution model of the telemedicine sensing big data database. The distribution structure of the medical sensing big data is shown in Fig. 1.

In the calculation, in order to ensure that the characteristic factors of telemedicine sensing big data are not disturbed, the characteristic factor quantity of the clustering center will use the fuzzy clustering factor ξ_i for clustering calculation. Through the fuzzy fusion correlation of data features, the fuzzy clustering function of data features in the database is obtained as follows:

$$c_i = c\left[x\left(y_0 + i\Phi y\right)\right] + \xi_i \tag{1}$$

Among them, c represents the calculation time in fuzzy clustering calculation; x indicates the amount of features;y represents a sequence node;$i\Phi y$ represents the data node;ξ_i represents the clustering factor.The fuzzy definition function of the feature distribution form of sensing big data is:

$$C_q(o) = \int_{-\infty}^{+\infty} c(y) r^{k\frac{y^2}{2}cot\beta - ktucsc\beta} dy \tag{2}$$

where, r represents the distribution order of data feature category and β represents the sampling quantity of characteristic distribution form; $\int$ represents the data distribution set;k represents a collection of classification properties.

To ensure uniformity of the calculated distribution, the definition standard of the data path characteristics is defined as follows:

$$K_V = \xi_i \sum_{n=0}^{i} \sum_{k=0}^{i} \beta(n,k) Q(n,k) \tag{3}$$

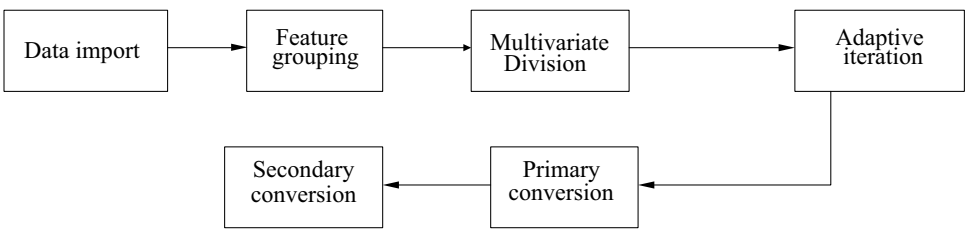

Fig. 1 The structure of medical sensing big data

In the formula,*n* represents the big data sequence analysis. By introducing the definition standard and normalizing the distribution form fuzzy function, the distribution model of data characteristics of a nonlinear database is obtained:

$$P = \sum_{i=0}^{I/2-1} 2\left(\xi_i cos\frac{2\pi li}{I} - sin\frac{2\pi li}{I}\right) \tag{4}$$

In the formula, πli represents ote the data edge node distribution.I represents the fitness operator.in which $l=0$, 1, …l-1. After that, the feature extraction of medical sensing big data is completed.

2.2 Telemedicine sensing big data denoising

Due to the different original environments of the extracted telemedicine sensing big data, a large amount of noisy data may lead to semantic drift, resulting in the highest utilization rate of medical sensing big data in this method. Therefore, it is necessary to perform noise reduction to the presence of seed relations in telemedicine sensing big data and train high confidence relational patterns. Let $f(t)$ represent the noisy signal that must be processed in the process of extracting feature-entity relationships in telemedicine sensing big data, and the expression is as follows:

$$f(t) = s(t) + n(t) \tag{5}$$

Among them, $n(t)$ describes the noise in the big data set, and $s(t)$ describes a valid signal. Given the noisy signal to be processed, the main purpose of the threshold denoising algorithm by Shearlet transform is to recover the effective seed relation quantity by Shearlet transform. The main steps of the Shearlet transform threshold denoising algorithm are as follows:

1. By the Shearlet transform, the noisy data of the sensing big data set is to obtain the Shearlet coefficient $C(j, l, k)$, which is expressed as follows:

$$C = f(t)C_q(o) \tag{6}$$

2. The signal coefficients were determined using the threshold function. $f(t)$ represents a minimum than threshold value, $C_q(o)$ value indicating that the coefficient is greater than the threshold value,Keep.The threshold setting procedure runs as follows:

$$C_{new}(j,l,k) = \begin{cases} C & |C(j,l,k)| \geqslant Th_j \\ 0 & |C(j,l,k)| < Th_j \end{cases} \tag{7}$$

Set N as the point of the denoising signal, j represents the dynamic feature division nodes; l represents ote the data distribution sequence; and the calculation formula Th_j for the threshold is as follows:

$$Th_j = \sigma\sqrt{2\ln(N)} \cdot \lg(j+1) \tag{8}$$

$$\sigma = \frac{median(|C(j,l,k)|)}{0.6745} \tag{9}$$

3. When the Shearlet coefficient $C_{new}(j, l, k)$ is processed by the threshold function, the inverse Shearlet transform is performed on the seed relation set to obtain the denoised telemedicine sensing big data $\hat{s}(t)$, and its expression is as follows:

$$\hat{s}(t) = P\sum_{j,l,k} C_{new}(j,l,k)\phi_{j,l,k} \tag{10}$$

$\Phi_{j,\,l,\,k}$ describes the threshold coefficient.

2.3 Telemedicine sensing big data feature fusion

Excessive model complexity reduces computational efficiency and model availability and is difficult to apply to real-world business scenarios. Deep learning technology can automatically learn sentence features during entity relationship extraction without the need for complex feature engineering. The problem of error propagation caused by language processing methods is avoided, and how can deep learning technology be combined to generate a highly integrated telemedicine sensing big data feature entity relationship extraction model to further improve the extraction effect of entity relationships is a feature entity of medical sensing big data, which is the key to relationship extraction. In this context, a feature fusion algorithm in deep learning is introduced to fuse the above extracted data features. n features are extracted from the big data of telemedicine class n to obtain the original data feature matrix, which is expressed as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} = \left[X_1, X_2, X_3, \cdots, X_p\right]^T \tag{11}$$

The original variables X_1, X_2, …, X_p can be used for linear representation of the comprehensive variables obtained after principal component analysis:

$$\begin{cases} Y_1 = U_1^T X = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1P}X_p \\ Y_2 = U_2^T X = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2P}X_p \\ \vdots \\ Y_p = U_p^T X = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pP}X_p \end{cases} \tag{12}$$

It is assumed that S_{ij} represents the covariance between the feature i and the feature j, and the calculation formula is as follows:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) \tag{13}$$

In the formula, x_{ik} represents a constraint model; $\bar{x}_i$ represents the vector feature distribution set; x_{jk} represents the dynamic fusion parameters; $\bar{x}_j$ represents the information fusion parameters.Construction of medical sensing big data fusion matrix according to the calculation results of the above formula S:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{14}$$

The eigenvalues are sorted from large to small to obtain each principal component. The eigenvalue of telemedicine sensing big data is the variance corresponding to each principal component, and $\lambda_1, \lambda_2, \cdots, \lambda_p$ is used to describe the non-zero eigenvalue corresponding to the eigenvector $U_1, U_2, \cdots, U_p$.

Let $\lambda = \lambda_k / \sum_{i=1}^{p} \lambda_i$ represent the contribution rate corresponding to the first principal component Y_k, describe the share of the information extracted by the k-th principal component in the total information, and obtain the cumulative contribution rate $\sum_{i=1}^{m} \lambda_i (\sum_{i=1}^{p} \lambda_i)^{-1}$ on this basis. The number of feature vectors and principal components is determined to obtain the transformation matrix. The transformation matrix is calculated using the principal component analysis and the original features to complete the feature fusion.

3 Feature entity relation extraction from telemedicine sensing big data

Based on the above fused medical sensing big data features, the samples are expanded and the similarity is calculated to determine the medical sensing big data feature entities, analyze the relationship between the entities and extract them.

3.1 Sample expansion

Due to the diversity of characteristics of each medical sensor data, there is a high similarity between the expanded data samples and the original samples. Therefore, by integrating the GAN model with the GMM model, the GMM-GAN data enhancement model is created and the model is used to realize the enhancement of medical sensing data and improve the accuracy of telemedicine sensing big data features.

When the sample of medical sensing data is expanded, it is more about understanding the medical sensing data feature entities, and the different medical sensing data features have different contributions to the analysis of the medical sensing data feature entity relationship. There is an interaction relationship to determine the features of medical sensing data, and medical sensing data pairs that meet the following conditions must be selected as candidates:

1. The distance between a pair of medical sensing data should be less than 3, i.e., only the pairs of medical sensing data in the same database and in only one database are considered.
2. For medical sensing data, the distance between features should be greater than 3 words and less than 50 words. The first condition is to set the distance between databases according to the pairs of features from medical data and to filter out many pairs of features from medical sensing data with too large distances. Usually, these pairs from medical sensing data have no interaction relationship. The second condition sets the required distance according to the distance between medical sensing data features. If the required distance is less than 50, the purpose is the same as the first condition, and the required distance is greater than 3. The purpose is to filter out the last two the features between the pair of candidate medical sensing data features and the left three words of the first medical sensing data feature entity and the right three words of the second medical sensing data feature form this pair of medical one candidate instance of the sensory data feature entity.

In this way, the GAN [10] model uses the generator G to make $q_{\text{data}}(G(z))$ closer to the sample distribution, and $q_{\text{data}}(G(z))$ represents the distribution of the sample $G(z)$. Thus, we can have $q_{\text{data}}(G(z), z)$. According to the multiplication formula of probability, the known prior distribution density functions $q_z(z)$ and $q_{\text{data}}(G(z), z)$ can be multiplied, and the formula is as follows:

$$\begin{aligned} q_{\text{data}}(G(z)) &= \int_z q(G(z), z)dz \\ &= \int_z q_{\text{data}}(G(z)|z)q_z(z)dz \end{aligned} \tag{15}$$

The diversity of $G(z)$ is reflected in the diversity of the prior distribution, q_z represents the coupling coefficient; dz represents the eigenvector,i.e., the generated samples may be more diverse due to the diversity of the prior distribution. If two entities in the database have a certain relationship, any text containing these two entities describes this relationship. This assumption is often not confirmed, resulting in a large amount of incorrect data in the generated database. So, in order to avoid the influence of this assumption on the

performance of relationship extraction, we assume that there is a pre-distribution density function of the GMM containing m components as $q_z(z)$, and that the covariance matrix of each Gaussian component is a diagonal matrix, which can be expressed as follows.

$$q_z(z) = \sum_{i=1}^{m} \pi_i N\left(z;\alpha_i, \beta_i\right) \tag{16}$$

Among them, $N\left(z;\alpha_i, \beta_i\right)$ and π_i are the probability density function and parameters of Gaussian mixture model, respectively. If there is too much noise, it is usually impossible to optimize the parameters π_i. Set $\pi_i = 1/m$, the formula is as follows:

$$N\left(x;\alpha_i, \beta_i\right) = \frac{1}{\left(2\pi^{(n/2)}\right)|\beta|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\alpha)^T \beta^{-1}(x-\alpha)} \tag{17}$$

In the formula, a represents a decision function, $N\left(z;\alpha_i, \beta_i\right)$ Select the repeated parameter adjustment technique to obtain the one-dimensional random noise vector subject to an priori distribution, and the formula is as follows:

$$z = \alpha_i + \beta_i \delta; \delta \sim N(0, 1) \tag{18}$$

where α_i represents the mean value of Gaussian components and β_i represents the standard deviation of Gaussian component. δ represents the maximum sampling threshold value.

The following formula can be obtained by combining Eqs. (14), (15) and (16):

$$q_{\text{data}}(G(z)) = \sum_{i=1}^{m} \int \frac{q_{\text{data}}\left(G\left(\alpha + \beta_i \delta\right)|\delta\right) q(\delta) d\delta}{m} \tag{19}$$

Among them, $\alpha = \left[\alpha_1, \alpha_2, \cdots, \alpha_N\right]^T$, $\beta = \left[\beta_1, \beta_2, \cdots, \beta_N\right]^T$, m and N represent the number of Gaussian components and the dimension of z, respectively.

Set the number interval of the Gaussian components to [20]. At this time, the redundant noise is fitted with the actual signal of mutual inductance [11], so that the sample effect obtained after data enhancement is the best. Add L_2 regularization term related to β to the loss function of generator G to avoid β value being 0. The corrected loss function formula of generator can be obtained as follows:

$$\min_G V_G(D, G) = \min_G E_{z\sim q_z}[\ln(1 - D(G(z)))] + \alpha \sum_{i=1}^{N} \frac{\left(1 - \beta_i\right)^2}{N} \tag{20}$$

The GMM-GAN model studied needs to initialize parameters. There are great differences in data distribution under different sample labels Y. It is necessary to initialize vectors α and β for each condition, so that $\alpha_i \sim U(-1, 1)$, $\beta \in (0, 1)$ and $U(-1, 1)$ are uniformly distributed in the interval (- 1,1), and the standard deviation can be randomly selected in the interval (0,1).

According to the above process initialization parameters, set $z = \alpha_k + \beta_k \delta, \delta \sim N(0, 1)$, take parameter k from 1 to m, input the obtained data into generator α_k to realize GAN model training, and train and optimize Gaussian component parameters α_k and β_k one by one through model training.

3.2 Calculation of sample similarity

Effectively filter the entity relationship collection composed of candidate medical sensor data features to improve the efficiency of entity relationship extraction. If there is no correlation between the data and the data, the text is represented by a vector to simplify the redundant relationship between the features of the medical sensor data. The database is regarded as composed of independent feature groups (T_1, T_2, …, T_n). Regarding each data T_i, a fixed weight W_i is given according to its criticality in the database, and (T_1, T_2, …, T_n) is regarded as a coordinate axis in an n dimensional coordinate system, (W_1, W_2, …, W_n). In order to compare the coordinate values, the database obtained by (T_1, T_2, …, T_n) decomposition is used in this way. Not only the frequency of feature items is considered, but the frequency of feature items is omitted, and the low-frequency weight is increased to ensure the effective extraction of entity relationships. Therefore, set a word frequency factor $\alpha\left(t_k\right)$, which is expressed as:

$$\alpha\left(t_k\right) = \frac{tf\left(t_k, C_i\right)}{\sum_{i=1}^{m} tf\left(t_k, C_i\right)} = \frac{\sum_{j=1}^{n} tf\left(t_k, d_{ij}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} f\left(t_k, d_{ij}\right)} \tag{21}$$

Among them, $tf\left(t_k, C_i\right)$ is the number of occurrences of feature items in the type, and $\sum_{i=1}^{m} tf\left(t_k, C_i\right)$ is the number of times feature item t_k appears in the database. The data training factor $\sum_{j=1}^{n} tf\left(t_k, d_{ij}\right)$ is the proportion of the medical sensor data including the feature item t_k in a specific class C_i occupying t_k in the entire database. The larger the $\alpha\left(t_k\right)$, the higher the frequency of the feature item in a specific class [12, 13], but the smaller the number of occurrences in other classes such a feature item has better ability to distinguish between types.

3.3 Entity relationship extraction

If the set of entity relations in the database is {A, B, C, …}, the probability of occurrence of a random type of an entity relation in U:

$$G(I) = \frac{\lambda_1}{\lambda_2} \tag{22}$$

where I represents the random item in the entity relationship set [14–16], λ_1 and λ_2 respectively represent the total number of such entity relationships that have been recorded in U. The weight value obtained by weighting the random items is:

$$\omega(I) = \frac{q}{G(I)} \tag{23}$$

where ω represents the weight; q indicates the applicability of entity relationship. Assume that the original time series is $T\{q * \tau\}$, it represents the number of time series data attributes [17, 18], and τ represents the number of timing data acquisition times. At this time, the symbolic processing for $T\{q * \tau\}$ includes:

$$H = \begin{pmatrix} H_{11} & \cdots & H_{1\omega} \\ \vdots & H_{ij} & \vdots \\ H_{q1} & \cdots & H_{q\omega} \end{pmatrix} \tag{24}$$

where H represents symbolic timing data, ω represents the number of data segments and H_{ij} represents the symbol pattern of the j-th attribute within $\left[i * \frac{\tau}{\omega}, (i+1) * \frac{\tau}{\omega}\right]$, where $\frac{\tau}{\omega}$ represents the compression rate of the segmented data. According to the symbol processing matrix [19, 20] shown in Eq. (24), time partition it and build the database within the time partition, let $i * \frac{\tau}{\omega} = T_1$ and $(i+1) * \frac{\tau}{\omega} = T_2$ correspond to the symbol sequence in column $\left[\iota_1, \iota_2\right]$ of (24), where, $\iota_1 = \frac{T_1}{\omega}$ and $\iota_2 = \frac{T_2}{\omega}$, substitute the above values into Eq. (24), and transpose Eq. (24), so as to obtain the database table with $\left|\iota_2 - \iota_1\right|$ rows and q columns:

$$H_T\left\{\left|\iota_2 - \iota_1\right|, q\right\} = \begin{pmatrix} H_{(\iota_1)1} & \cdots & H_{(\iota_1)q} \\ \vdots & H_{(\iota_j)i} & \vdots \\ H_{(\iota_2)1} & \cdots & H_{(\iota_2)q} \end{pmatrix} \tag{25}$$

where t_j represents the j-th attribute time partition, and $H(t_j)_i$ represents the symbol pattern of the i-th attribute in $\left[\iota_j * \frac{\tau}{\omega}, \left(\iota_j + 1\right) * \frac{\tau}{\omega}\right]$.

Based on the above calculation process, the frequent itemset tree is used to generate the data frequent itemset, and the database $T\{q * \tau\}$ is traversed, then:

$$H_{T-h} = \begin{pmatrix} h_{1\_(\iota_1)1} & \cdots & h_{1\_(\iota_1)q} \\ \vdots & h_{g\_(\iota_j)j} & \vdots \\ h_{\left|\iota_2-\iota_1\right|\_(\iota_2)1} & \cdots & h_{\left|\iota_2-\iota_1\right|\_(\iota_2)q} \end{pmatrix} \tag{26}$$

$H_{T\text{-}h}$ represents the data frequent itemset matrix, and $h_{g(\iota_j)j}$ represents the data item traversed for the second time. According to the frequent itemset matrix, judge whether $h_{g\_(\iota_j)i}$ exists in the row of Eq. (25), then:

$$h_{g\_(\iota_j)j} = \begin{cases} 1, \delta \in H_{\iota_j} \\ 0, \delta \notin H_{\iota_j} \end{cases} \tag{27}$$

where H_{ι_j} represents the symbol mode; Indicates presence; Indicates that it does not exist. The column vector count $\sum_{g=1}^{\left|\iota_2-\iota_1\right|} h_{g\_(\iota_j)j}$ in H_{T-h} and judge whether it meets the following conditions:

$$\sum_{g=1}^{\left|\iota_2-\iota_1\right|} h_{g\_(\iota_j)j} \geqslant \varepsilon \tag{28}$$

where ε represents the minimum limit. If $\sum_{g=1}^{\left|\iota_2-\iota_1\right|} h_{g\_(\iota_j)j}$ meets the conditions shown in Eq. (12), the frequent itemset tree is constructed according to the data frequent itemset matrix shown in Eq. (28). At this point, the established frequent itemset tree is the association rule for medical information time series. The weighted processing of medical information, telemedicine and sensing big data, and the weighting of telemedicine and sensing big data as ωt (U), exist:

$$\omega t(U) = f \sum_{i=1}^{n} \frac{\omega_i(I)}{|U|} \tag{29}$$

where f represents the empirical value, ω_i indicates the number of queries, and $|U|$ represents the total number of information sets in the database that meets the requirements. The above content is fused with Eqs. (23) and (24) to obtain:

$$\begin{cases} \omega - support(A \Rightarrow B) = \sum_{i=1}^{n} \omega(I) / \sum_{i=1}^{m} \omega(I) \\ \omega - confidence(A \Rightarrow B) = \sum_{i=1}^{n} \omega_i(U) / \sum_{i=1}^{m} \omega_i(U) \end{cases} \tag{30}$$

where, m represents the feedback times of the database. Through the fusion process described above, the results of entity relationship division and association rules are fused, and the feature entity relationship of telemedicine sensor big data is extracted according to the fusion process. According to the concept of information matching, the feature entity relationship and association rules are used to match information. The calculation formula is:

$$\mu(\sigma = 1|\alpha, \beta) = \sum_{\alpha} \sum_{\beta} \sum_{\varphi} \sum_{\eta} \mu(\sigma = 1, R, F, O, K|\alpha, \beta) \tag{31}$$

where σ is the two characteristic correlation parameters obtained by PSO according to Eq. (28). $\mu(\sigma = 1, R, F, O, K|\alpha, \beta)$ represents the feature extraction function. When the value is set to 1, it is proved that the relationship between the two entities has strong correlation, and φ and η are the possible matrices of O and K.

4 Experiment

4.1 Experimental data and process

To verify whether the deep learning-based method for extracting feature-entity relationships from big data in telemedicine can effectively extract the data-feature-entity relationship, 10 representative relationship labels were selected from the 96 known relationship types in the YAGO database. experiment. Some of these 10 relationship labels can express a variety of characteristic entity relationships, and some contain entity pairs with multiple relationship labels in the database, which can easily lead to a noise labeling problem. Since some databases in the Wikipedia corpus are complex and too long, this chapter also filters out sentence instances with sentence lengths of less than 3 and more than 80. The parameters of the model GMM-GAN are set, i.e., the pre-trained database is used for initialization and the English Wikipedia is trained using Google's open-source word2vec tool. The dimension of the word data is set to 50, the dimension of the position vector is set to 5, the dimension of the data feature vector is set to 230, the dimension of the relation vector is set to 230, the learning rate is set to 0.001, the batch size is set to 60, and the dropout rate is set to 0.5.

4.2 Experimental indexes and results

(1) Similarity judgment

In the medical sensing database included in the above 10 relationship types, they are divided into 10 groups according to the data relationship types, and each group manually selects 100 data features similar to the corresponding relationship labels as experimental data, and then calculates the shortest dependency path and relationship similarity between each data feature entity using different extraction methods [2], and finally takes the average of the 100 similarity results for each relationship type as the final calculation result. The results of the similarity comparison are shown in Fig. 2.

In Fig. 2, the similarity of entity relations extracted by this method is more than 70% compared to other methods. Thus, the semantic similarity between phrases and relation phrases can be accurately measured on the shortest dependency path to assess whether the labeled data is correctly labeled.

(2) Extraction time

The extraction time of the different methods [2, 3] is analyzed based on the different times for matching association rules, and the result of the extraction time comparison is shown in Fig. 3.

According to Fig. 3, the extraction time of the method in this paper is the shortest and the volatility is low. Combined with deep learning technology, a telemedicine information relationship extraction model with a high degree of integration is built, namely the data enrichment model GMM-GAN. For the medical sensor big data with interactive relationships in the sensor database, the

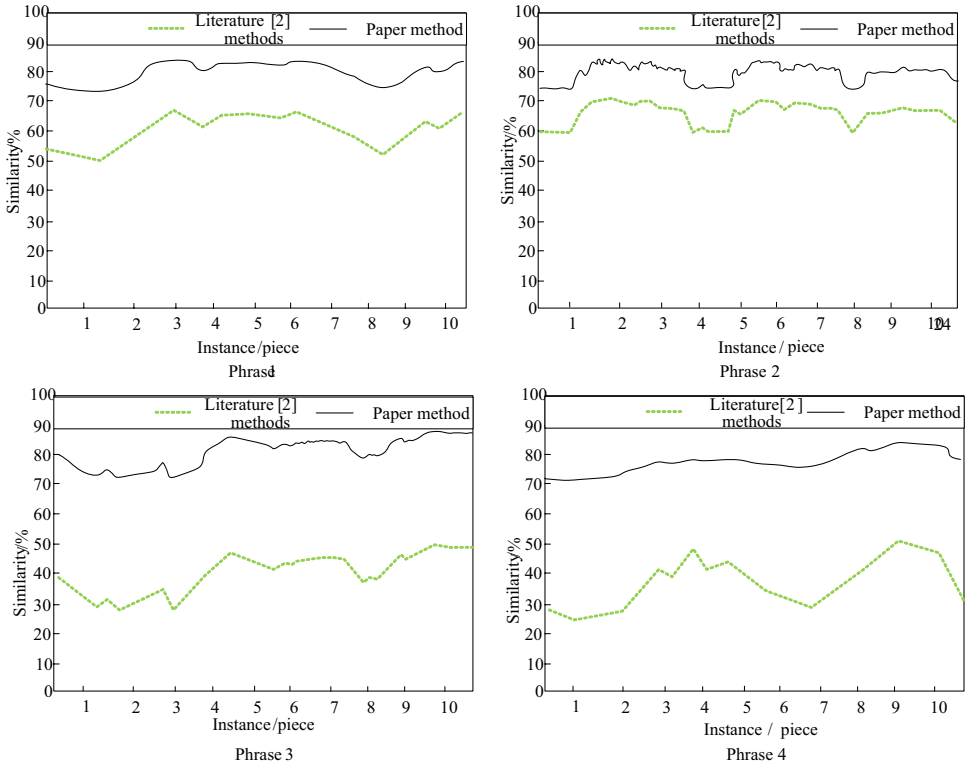

Fig. 2 Comparison results of similarity

association rules are matched one by one. If the matching is successful, the extraction starts directly, which reduces the amount of intermediate data and improves the extraction efficiency.

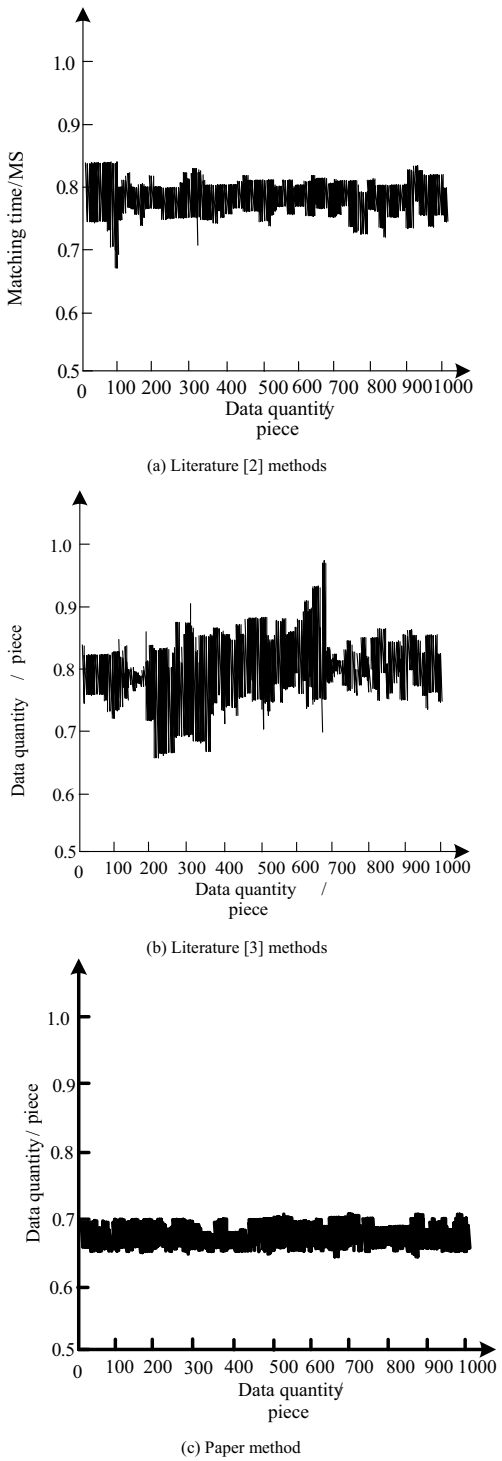

Fig. 3 Comparison results of extraction time

5 Conclusion

In this paper, combined with deep learning technology, we build a complete entity relationship extraction model from telemedicine sensor data, analyze whether there is depth relationship between entities, and calculate the transformation matrix through principal component analysis to complete the original features. Then, we select high-quality seed relation sets, train high confidence relation patterns for noise reduction, match association rules with noise free data structures, and extract medical sensor data pairs with interactive relationships. By obtaining the adaptive feature clustering center, setting the probability density function and parameters of Gaussian mixture model, and constraining the feature items with the highest similarity, the data features in the database around the feature clustering center will reconstruct the distribution of similar features, such as the information text of different types of telemedicine video equipment and precision medical related equipment. The results show that the developed model has better performance than other methods. The experimental results show that the similarity of entity relationship of this method is more than 70%, and the extraction time of this method is less than 0.7, which shows the effectiveness of this method. In the future research, we need to do more in-depth research on the feature entity relationship extraction method of medical perception big data.

Funding Open access funding provided by Western Norway University Of Applied Sciences

Declarations

Conflict of Interest The authors have no relevant financial or non-financial interests to disclose. Wenkui Zheng provided the algorithm and wrote the original manuscript, Wei Hou did the experiment and analyzed the experimental results, Jerry Chun-Wei Lin revised the paper, supervised and analyzed the experiment. We also declare that data availability and ethics approval is not applicable in this paper.

References

1. Yuyong C, Xinghua L (2021) Big data anomaly extraction algorithm based on uncorrelation test. Computer simulation 38(03):245-248+460
2. Antonino G, Lorenzo C, Alessia B, Maria F (2019) An innovative methodology for big data visualization for telemedicine. IEEE Trans Industr Inf 15(1):490–497
3. Jiazhen Z, Hongbo S, Bing S, Shuai T, Yang T (2020) Deep neural network based recursive feature learning for nonlinear dynamic process monitoring. Canadian J Chem Eng 98(4):919–933
4. Wang S-H, Nayak DR, Guttery DS, Zhang X, Zhang Y-D (2021) COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. Information Fusion 68:131–148
5. Wang M, Lin Y, Tian Q, Si G (2021) Transfer learning promotes 6G wireless communications: recent advances and future challenges. IEEE Trans Reliab 70(2):790–807
6. Gao P, Li J, Liu S (2021) An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. Mobile Networks & Applications 26(5):2123–2126
7. Wang S-H, Govindaraj VV, Gorriz JM, Zhang X, Zhang Y-D (2021) Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network. Information Fusion 67:208–229
8. Mitchell M, Kan L (2019) digital technology and the future of health systems. Health Syst Reform 5(2):113–120
9. Zhong G, Ling X, Wang L-N (2019) From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. Wiley Interdiscip Rev Data Min Knowl Discov 9(1):251–258
10. Zirui L, Xueqing T, Tianjia G et al (2021) Third-party institutions participate in healthcare big data sharing mode [J]. Public Health China 37(7):4
11. Qiang G, Cong W (2021) Opportunities, challenges and their development of medical big data platform construction [J]. J Med Inform 42(1):7
12. Niousha BK, Mohammad SA, Saeed S-A, Sadat JN (2021) Cancer miRNA biomarkers classification using a new representation algorithm and evolutionary deep learning. Soft Comput 25(4):3113–3129
13. Wang Qiangfen Lu, Fenghua CS et al (2020) Investigation and analysis of medical students' information literacy and its influence in the Era of Medical Big Data [J]. Health Serv Manag China 37(2):6
14. Liu S, Wang S, Liu X, Lin C-T, Lv Z (2021) Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans Fuzzy Syst 29(1):90–102
15. Wang S, Celebi ME, Zhang Y-D, Yu X, Lu S, Yao X, Zhou Q, Miguel M-G, Tian Y, Gorriz JM, Tyukin I (2021) Advances in data preprocessing for biomedical data fusion: an overview of the methods challenges and prospects. Information Fusion 76:376–421
16. Jujie W, Xin S, Qian C, Quan C (2021) An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting. Sci Total Environ 762(8):143–149
17. Sakkari M, Zaied M (2020) A Convolutional Deep Self-Organizing Map Feature extraction for machine learning. Multimed Tools Appl 79(27–28):19451–19470
18. Hoyle P (2019) Health information is central to changes in healthcare: A clinician's view. HIM J 48(1):48–51
19. Liu S, Wang S, Liu X, Gandomi AH, Daneshmand M, Muhammad K, De Albuquerque VHC (2021) Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans Multimed 23:2188–2198
20. Liu S, Liu X, Wang S, Muhammad K (2021) Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT assisted complex environment. Neural Comput Appl 33(4):1055–1065