ATLANTIS PRESS

Advances in Engineering Research, volume 166

3rd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2018)

Study on Text Mining Algorithm for Ultrasound Examination of Chronic Liver Diseases Based on Spectral Clustering

Bingguo Chang[a, *], Xiaofei Chen[b]
School of Information Science and Engineering, Hunan University, Hunan 410082, China;
a531980691@qq.com, b735220914@qq.com

***Abstract*—Ultrasonography is an important examination for the diagnosis of chronic liver disease. The doctor gives the liver indicators and suggests the patient's condition according to the description of ultrasound report. With the rapid increase in the amount of data of ultrasound report, the workload of professional physician to manually distinguish ultrasound results significantly increases. In this paper, we use the spectral clustering method to cluster analysis of the description of the ultrasound report, and automatically generate the ultrasonic diagnostic diagnosis by machine learning. 110 groups ultrasound examination report of chronic liver disease were selected as test samples in this experiment, and the results were validated by spectral clustering and compared with k-means clustering algorithm. The results show that the accuracy of spectral clustering is 92.73%, which is higher than that of k-means clustering algorithm, which provides a powerful ultrasound-assisted diagnosis for patients with chronic liver disease.** *(Abstract)*

Keywords-component; Text mining; Ultrasonic examination report; Spectral clustering learning algorithm; Chronic liver disease (key words)

I. INTRODUCTION

Ultrasound examination is to observe the body by the reflection of the human body for ultrasound .It irradiates a weak ultrasound on the body. Then the reflected wave of the examination parts is processed into image. Such operations in diagnosis of chronic liver disease is widely used. Machine learning can quickly obtain ultrasound results, and can eliminate harmful interference factors.

In recent years, significant achievements have been made in biomedical text processing technologies. Using biomedical text mining tools to automatically extract the frequency of rheumatic symptoms [2]. Health care analysis system adopts an approach, which is based on dictionary segmentation and use iterative feedback mechanism to enhance self-learning [3]. The text feature classification, which is based on Membrane Particle Swarm Optimization and Information Entropy, can significantly improve the classification accuracy[4] .Data mining, which is based on latent semantic tree analysis model ,is used for medical texts.[5] Multimodal Semi - supervised Learning Model Based Multi - modal Information Retrieval Algorithm for Similar Case Finding[6].The similarity analysis of medical texts, which is based on the full-text indexing and cosine formula, analyzes the similarity of medical texts[7]. Aiming at the problems of synonym recognition and naming standardization of disease diagnostic text in electronic medical records, an adaptive text clustering method was raised [8]. Based on the characteristics of report texts in ultrasound examination, this paper presents a method of using spectral clustering to automatically generate examination prompts through machine learning.

Based on the characteristics of report texts, this paper presents a method of using spectral clustering to cluster ultrasound reports according to the ultrasound findings. The ultrasound diagnosis in the same category is selected by physicians to assist doctors in making ultrasound diagnosis.

II. MINING FOR ULTRASOUND EXAMINATION TEXT

A. the source of Experimental data

Based on the sample datasets provided by the Big Data Project of the Third Xiangya Hospital of Central South University, this paper selects 110 cases of ultrasound examination report from March 2008 to May 2016 in the Third Xiangya Hospital. These patients were eventually diagnosed with chronic liver disease of varying severity.

B. experimental method

Ultrasound examination report text mining is divided into three steps: The first step, which is called the segmentation processing, the description part of the ultrasound examination report text is segmented for the sake of the formation of the corpus. In the second step, a document term relationship matrix is generated by calculating the value of TF-IDF (term frequency-inverse document frequency). In the third step, Gaussian kernel algorithm is proposed to obtain the similarity matrix. Laplace matrix is calculated, and the first 10 eigenvectors of the Laplacian matrix are obtained as 10 center values to cluster the ultrasound examination report texts.

C. the processing of word segmentation

Based on the Trie-tree structure, an efficient word map scanning is realized to generate a directed acyclic graph (DAG) composed of all possible Chinese words in a sentence. The dynamic programming is used to find the maximum probability path to find the maximum segmentation combination based on word frequency For unregistered words, the HMM model based on Chinese word formation ability is used, and the Viterbi algorithm is utilized. We use the Tinn-R software to analyze the text of the description of the ultrasound report texts utilizing the Jieba algorithm. Partial segmentation results as follows:

卧位 扫查 左肝 右肝 斜径 形态 规则 轮廓 清 实质 回声 近
场 增强 远场 衰减 光点 细密 肝内 管系 结构 显示 欠清 肝
内 胆管 未见 扩张 内未见 结石 声像 肝内 多个 无 回声 区
界清 有 包膜 较大 者 位于 左肝 外叶 大小 为 胆囊 大小 壁
光滑 连续 内透声 可 未见 结石 声像 门静脉 内径 内 清晰
胆总管 内径 内 清晰 胰头 胰 体 胰 尾 大小 为 形态 规则
轮廓 清 实质 回声 均匀 主 胰管 未见 扩张 未见 肿块 声像
脾厚 形态 规则 轮廓 清 实质 回声 均匀 未见 肿块 声像 左
肾 大小 右肾 大小 形态 规则 轮廓 清 实质 回声 低于 肝 脾
双肾 集合 系统 未见 分离 未见 肿块 结石 声像 双 输尿管
未见 扩张 膀胱 充盈 好 内壁 光滑 透声 好 未见 肿块 结石
声像 腹腔 未见 液 暗区 腹膜 后 未 见 肿大 淋巴结 声像 经
腹 扫查 前列腺 大小 约 形态 规则 轮廓 清 实质 回声 均匀
实质 内未见 肿块 声像 门静脉 血流 充填 好 双肾 血流 信号
呈 树枝状 分布 各 脏器 未 探及 异常 血流 信号

Figure 1. Word segmentation results

D. Generate document-term matrix

Definition 1: Suppose A represents the document entry matrix, the rows represent documents, and the columns represent entries. The element of matrix A refers to the TF-IDF value, the value of which is the product of Term Frequency (TF) and inverse document frequency (IDF) of the j-th entry of the i-th document. Because the word frequency value of a lot of entries is zero, the document-term matrix is sparse matrix.

Definition 2: It is assumed that d documents consist of m terms. a_{ij} is the element point of the document-term matrix. t_{ij} is the number of occurrences of the j-th term in document i, that is, the term frequency (TF); d_j is j terms of the number of documents. There is

$$a_{ij} = \frac{t_{ij}}{\sqrt{\sum_{j=1}^{n} t_{ij}^2}} \lg \frac{d}{d_j} \quad (1)$$

The weight of words is described by the product of word frequency TF and reverse word frequency IDF. According to formula (1), the high-frequency words in a particular file and the low-frequency words in the entire file set will generate TF-IDF values of high weights, while the weights of commonly used words that have little effect on the clustering will be given smaller value.

For example: In this experiment, "missing" may appear in each case of ultrasound examination report, the frequency is very high, so there is not much to the clustering. By using formula 1, the weight of the word is zero, which will not affect the accuracy of the entire cluster.

E. analysis and description of spectral clustering algorithm

The spectral clustering algorithm treats data clustering as a multi-partitioning problem of directional graphs, and considers each text as the vertex of the graph. The connecting edges between the vertices are weighted by the similarity between the texts for the obtaining of a non-directional weighted graph based on text similarity .Following the principle of optimal segmentation--the similarity between any two subgraphs is the smallest and the similarity within the subgraph is the largest. The graph segmentation problem is transformed into the spectral decomposition of the Laplacian matrix to get the global optimal solution of the graph segmentation in the continuous domain.

The spectral clustering algorithm mainly consists of the following steps:

1. Calculate the similarity matrix of data samples, and predefine number of clusters.

Similarity matrix, also known as the affinity matrix, the kernel function is generally used to obtain the similarity matrix. By comparing various kernel functions, it is found that the clustering effect of Gaussian kernel function is the best. Therefore, the Gaussian kernel function algorithm is chosen to calculate the similarity matrix. Gaussian kernel function is calculated as follows:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (2)$$

Among them, x, x 'represents the sample points of the data, that is, any two row vectors in the document-term matrix A. $\|x - x'\|$ is the Euclidean distance between the sample points x and x ', and the parameter specified by σ determines the attenuation rate between data points. In this experiment, σ = 1 is specified.

2. Calculate Laplace matrix

Definition 3: The sample set is represented by an undirected graph G = (V, E). The vertex v_i in V= $\{v_1, v_2, \ldots, v_n\}$ represents the data point in the sample set and E represents the corresponding edge. The weight of the edge of sample i and sample j is equal to the element W_{ij} in the similarity matrix, that is, the undirected graph G = (V, E) represents the similarity matrix. Given n data samples $x_1, x_2, \ldots, x_n$, the similarity of the samples xi and x_j is expressed by $W_{ij} \geqq 0$. When xi is similar to x_j, W_{ij} is represented by a larger value; when xi is not similar to x_j, W_{ij} is represented by a smaller value. The similarity is symmetrical, that is, we can assume $W_{ij} = W_{ji}$. Where for any i, $W_{ii} = 1$.

D = diag ($dg_1, dg_2, \ldots, dg_n$), where dg_i is the sum of the i-th row of elements in the similarity matrix.

From the above we can define the Laplacian matrix:

$$L = D - W \quad (3)$$

Regularized Laplacian matrix

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \quad (4)$$

3. Find the first k largest eigenvalues of the matrix L_{sym} and the corresponding eigenvectors $v_1, v_2, \ldots, v_k$, construct the eigenvector space matrix $V=\{v_1, v_2, \ldots, v_k\} \in R^{n*k}$.

Solve the eigenvectors of the matrix L_{sym}, V represents the eigenvectors, λ expressed as eigenvalues. Use the following formula:

$$(I - D^{-1/2} W D^{-1/2}) V - \lambda V = 0 \quad (5)$$

4. Each row of the eigenvector space matrix is regarded as a point in space, and it is clustered into k classes by K-means clustering method.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental results

According to the above step 1, in the case of clustering number k = 10, in order to verify the effect of the spectral clustering algorithm in the text extraction of the ultrasound examination report, 110 sets of data sets of the chronic liver disease ultrasound examination report text were used for experiments and tests. TF-IDF method was used to extract the features of 110 ultrasound examination texts. In order to reduce the sparseness of matrix A, the number of Word frequency value which is less than 5 were excluded. According to formula (1), the result of document-term matrix A is as follows:

$$A_{110*326} = \begin{pmatrix} 0.00053 & 0.0039 & 0.00261 \\ 0.00079 \cdots & 0 & \cdots 0.00386 \\ 0.00113 & 0.00415 & 0.00275 \\ \vdots & \cdots & \vdots \\ 0.00075 & 0 & 0.00367 \\ 0.00069 \cdots & 0 & \cdots \quad 0 \\ 0.00070 & 0.00517 & 0.00342 \end{pmatrix}$$

According to formula (2) Calculate the similarity matrix W of 110 experimental texts:

$$W_{110*110}=\begin{pmatrix} 1 & & 0.9670 & & 0.9829 \\ 0.9909 & \cdots & 0.9680 & \cdots & 0.9862 \\ 0.9820 & & 0.9682 & & 0.9767 \\ \vdots & & \ddots & & \cdots \\ 0.9856 & & 0.9658 & & 0.9839 \\ 0.9849 & \cdots & 0.9657 & \cdots & 0.9799 \\ 0.9829 & & 0.9663 & & 1 \end{pmatrix}$$

When the image characteristics in Ultrasound examination report text is described, due to other reasons such as doctor's personal habits, most of the text reports follow a certain format, and the similarity between the texts has reached more than 0.96. For example, most reports of chronic liver disease on ultrasound will first describe the shape and size of the left liver and liver, as well as some information about the gallbladder, pancreas and other organs, which are mostly similar. In addition, because this article studies the ultrasound examination report text in the special field of chronic liver disease, the related terms are relatively fixed. Due to the limitations of the scope of the report text of ultrasound examination of chronic liver disease, it will not be like other large-scale text clustering which there is a big difference.

The Laplace matrix is calculated from equations (3) and (4) and the result is as follows:

$$L_{sym110*110}=\begin{pmatrix} 0.00935 & & 0.00918 & & 0.00921 \\ 0.00925 & \cdots & 0.00918 & \cdots & 0.00922 \\ 0.00920 & & 0.00922 & & 0.00917 \\ \vdots & & \ddots & & \vdots \\ 0.00921 & & 0.00918 & & 0.00921 \\ 0.00921 & \cdots & 0.00917 & \cdots & 0.00918 \\ 0.00921 & & 0.00922 & & 0.00939 \end{pmatrix}$$

According to the formula (5), the eigenvalues of the matrix L are calculated. The eigenvectors corresponding to a maximum of 10 eigenvalues are selected and the eigenvector space matrix is formed:

$$V_{110*326}=\begin{pmatrix} 0.00077 & & 0.00332 & & 0 \\ 0.00062 & \cdots & 0.00165 & \cdots & 0 \\ 0.00271 & & 0.00158 & & 0.02371 \\ \vdots & & \ddots & & \vdots \\ 0.00905 & & 0 & & 0 \\ 0.00076 & \cdots & 0.00613 & \cdots & 0 \\ 0.00044 & & 0.00646 & & 0 \end{pmatrix}$$

Based on the 10 eigenvectors obtained by the spectral clustering method described in the paper as the central value of the K-means algorithm, 110 test samples are grouped into 10 categories.

Because it is difficult to describe the distribution of data in 326 lexical space. This article selected the "depth", "chest", "free" in the three groups of 110 data distribution of the highest frequency of entry, in pairs as a group, respectively, draw the distribution of 110 sets of data in these two terms. Each set of data is a point on the plane. When two points coincide, only the color of the next point will be displayed. The figure below shows the case where the three terms are clustered respectively on the x-axis and the y-axis. Different color points are on behalf of different classes. The color of the first to the tenth class, respectively, with red, green, blue, gray, tan, lavender, khaki, coral, purple, gold .As seen in Figure 2.

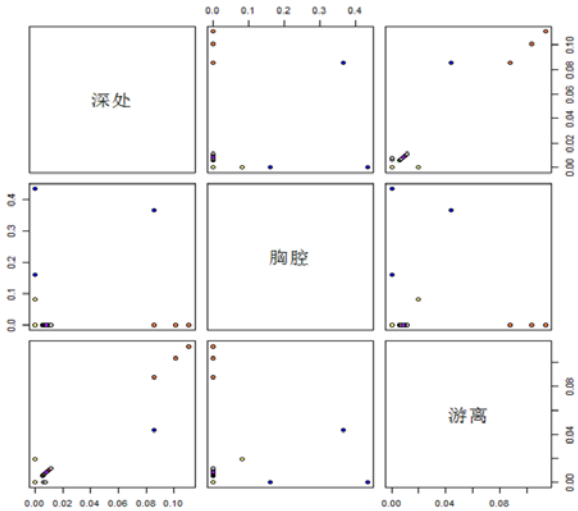

Figure 2. The corresponding two-dimensional plane between the different terms of the cluster situation

According to the clustered situation, some of the 10 most frequently occurring words are displayed in the form of word clouds. The higher the frequency of words is, the larger the font size is. The first in the first line is the first category of clustering, then the first of the second line is the second category, and so on. As shown in Figure 3:

Figure 3. 9 types of high frequency word cloud

By looking at Figure 2, it is basically possible to determine the higher frequent and specific terms for each category. For example, the most frequent words in category c_1 are "vein", "blood flow", and so on. However, for both categories c_6 and c_{10} (i.e. the second and third columns of the third row), the high-frequency terms for both categories are "medium", where category c_{10}, "far Field "," fine "and other terms compared to the type of c_6 are higher, there is still some difference between the two. In addition, there is no indication due to the small number of entries in category c_8. The above results can help doctors summarize a class of chronic liver disease ultrasound report to provide assistance. For example, category c_3, which has the highest frequency of occurrence, is the term "thoracic cavity", and it is assumed that the ultrasound report of the type is most closely associated with the thorax. When the analysis of such illness summary in the future, it can be a good analysis of this link.

B. quality evaluation of clustering

In order to estimate the clustering quality of the proposed algorithm, we used the purity calculation precision, the set C = {1, 2, ..., m} is obtained by artificially adding the classification identifier set g = {1, 2,…, m}, Then is defined as the number of documents that belong to cluster c_j in classification identifier i. n_j is defined as the number of documents that form the c_j cluster.

Purity p is defined as

$$p = \frac{1}{d}\sum_{i=1}^{m} \max_j \{n_j^i\} \quad (5)$$

Where i is a class identifier variable and d is the total number of documents, $\max_j \{n_j^i\}$ is the maximum number of documents in the cluster set C which belongs to the classification identifier i.

This paper compares the k-means clustering algorithm in the ultrasound text on the clustering results. Then to base on the formula (5) calculate the purity, the result is as follows:

Table 1 Comparison of two clustering algorithms in text mining

Comparison of two clustering algorithms	**Property**			
	stability	***Sample distribution***	***Computational complexity***	***Accuracy***
Spectral clustering	stable	Any	Simple	92.73%
K-means	Unstable	Special	Complex	83.64%

As can be seen from Table 1, the spectral clustering algorithm has higher purity in processing ultrasound texts than the direct k-means algorithm. The spectral clustering algorithm calculates the similarity matrix between texts through kernel function, and it is significant for the clustering effect of processing sparse sample features and reduces the computational complexity of the matrix. Compared with the direct use of k-means algorithm, spectral clustering has the ability to cluster any shape of sample space and converge to the global optimal solution in terms of sample shape distribution. In terms of stability, the clustering results obtained by spectral clustering on each basis of the selected kernel function have good stability.

IV. CONCLUSION

The World Health Organization's reports that about 400 million people worldwide are suffering from chronic liver disease, which is a potentially harmful and widespread epidemic with a low cure rate and high mortality. Ultrasound is an important diagnostic item for chronic liver disease. In this paper, the method of spectral clustering is used to cluster the ultrasound examination reports according to the ultrasound findings. The ultrasound diagnosis in the same category is selected to assist the doctor with the ultrasound diagnosis, thereby reducing the workload of doctor. In this paper, 110 groups of chronic liver disease ultrasound examination report texts were selected as samples, which were verified by spectral clustering and compared with the direct use of k-means clustering algorithm. The results show that the accuracy of k-means clustering algorithm is 83.64%, while the accuracy of spectral clustering algorithm is 92.73%, which shows that the spectral clustering algorithm in this paper is more effective and provides a powerful technical support for ultrasound-assisted diagnosis. However, because of the nature of the report text itself in the ultrasound examination of chronic liver disease, most of the texts are highly similar in terms of similarity. Therefore, how to better reduce the text similarity between different categories is also a focus of later research.

V. REFERENCES

[1] Yan Xu.Application of Machine Learning Technology in Medical Data Mining[J].Tianjin: China CIO News ,2016(1):89-89.

[2] P Yildirim,Çinar Çeken,R Hassanpour, et al.Prediction of Similarities Among Rheumatic Diseases[J] , Netherlands:Journal of Medical Systems, 2010, 36(3):1485-90

[3] Zheng Jye Ling, Quoc Trung Tran, Ju Fan et al.GEMINI: An Integrative Healthcare Analytics System[J]. Proceedings of the Vldb Endowment, 2014, 7(13):1766-1771

[4] DOU Zengfa,GAO Lin. Feature Selection of Biomedical Literature by Membrane Particle Swarm Optimizer and Information Entropy.Xian: Journal of Xi'an Jiaotong University, 2012, 46(4)：0253-987X.

[5] LI Bo,WEN Dunwei,WANG Ke et al.Automatic annotation for medical texts based on hidden topic and semantic tree[J].Changchun:Journa of Jilin University Engineering and Technology Edition ,2012,42(1)：1671－5497.

[6] Wu Menglin,Chen Qiang,and Sun Quansen.Medical Case Retrieval Based on Combination of Images and Textual Information[J].Beijing: Journal of Computer-Aided Design and Computer Graphics,2014,26(9):1430-1437.

[7] Xie Cuiping, Chen Jiayi, Bai Jinshan. Similarity Analysis of Medical Text Based on Full-Text Indexing Technology and Cosine Formul.Shanghai: Microcomputer Applications ,2014, 30(1):25-27.

[8] Li Wei, Xu Hongtao, Zhao Dazhe et al. An AdaptiVe Clustering Method on Medical Short Text[J] .Shenyang:Journal of Northeastern University :Natural Science, 2015,36(1):19-23.