Research Article

Weipeng Zhang*

Relationship between solitary pulmonary nodule lung cancer and CT image features based on gradual clustering

DOI 10.1515/phys-2017-0044
Received July 28, 2016; accepted September 6, 2016

Abstract: Background: The relationship between the medical characteristics of lung cancers and computer tomography (CT) images are explored so as to improve the early diagnosis rate of lung cancers. Methods: This research collected CT images of patients with solitary pulmonary nodule lung cancer, and used gradual clustering methodology to classify them. Preliminary classifications were made, followed by continuous modification and iteration to determine the optimal condensation point, until iteration stability was achieved. Reasonable classification results were obtained. Results: the clustering results fell into 3 categories. The first type of patients was mostly female, with ages between 50 and 65 years. CT images of solitary pulmonary nodule lung cancer for this group contain complete lobulation and burr, with pleural indentation; The second type of patients was mostly male with ages between 50 and 80 years. CT images of solitary pulmonary nodule lung cancer for this group contain complete lobulation and burr, but with no pleural indentation; The third type of patients was also mostly male with ages between 50 and 80 years. CT images for this group showed no abnormalities. Conclusions: the application of gradual clustering methodology can scientifically classify CT image features of patients with lung cancer in the initial lesion stage. These findings provide the basis for early detection and treatment of malignant lesions in patients with lung cancer.

Keywords: Solitary pulmonary nodule, CT image, gradual clustering

PACS: 87.85.-d

1 Introduction

Lung cancer is the malignant tumor with highest morbidity and mortality in the world. The formal expression of most lung cancers are solitary pulmonary nodule in the early stage [1, 2]. These nodules are usually quasi-circular lesions with diameters less than 3 cm, which have different shapes, indefinite distribution locations and easy adhesion with other organizations [3]. Clinical symptoms do not usually exist, and the features are not very obvious on the CT image, therefore, determining whether the nodule is benign or malignant is a key to prevention of secondary lung cancer [4, 5]. In automatic identification of a solitary pulmonary nodule, understanding the feature extraction and expression of pulmonary nodule in CT images are the key point.

This paper takes advantages of methods based on fuzzy clustering of medical signs of early-stage lung cancer; using this combination, we hope to achieve early detection and timely treatment of malignant lesions. Frequently-used fuzzy clustering methods include the dynamic clustering method, systemic clustering method and fuzzy C-mean value algorithm. This paper adopts dynamic clustering, which is also called the gradual clustering algorithm of fuzzy clustering. Gradual clustering has the advantages of small computational effort, small computer memory space requirements, and flexibility.

At present, as an important technology in data mining, gradual clustering is widely used in medicine. Gradual clustering analysis is often used to provide methodology for classification for traditional Chinese medicine (TCM) clinic treatment based on syndrome differentiation [6]. For example, Zhang Mingxue, et al. applied statistical methods such as clustering analysis to results and summed up the types and characteristics of syndromes in four phases of coronary heart disease complicated with hypertension [7]. Dynamic clustering analysis is also used for studies related to medication rules and screening of medication regimens in prescriptions, etc. Zhou Lu et al. applied a clustering analysis method in fuzzy mathemat-
ics to discuss the mutual relations between the TCMs relieving an exterior syndrome [8]. Fuzzy dynamic clustering analysis is also applied in medical image processing; for example, Tian Jie et al. applied this method to three-dimensional medical-image processing and analysis, including CT, spiral CT and MRI, thus better identifying thin bones and bones at the articulated joints; after rebuilding, the 3D model can clearly reproduce the anatomical structure [9].

2 Research methods

2.1 Inspection method

CT scans were executed on patients separately. The CT scanner with 64 rows of spiral produced by USA GE Healthcare has been adopted to allow inspection and diagnosis on patients. Through a mediastinum window, the scans allow measurement of the maximum diameter of the lesion and evaluation of the lobulated calcification condition. The window width was set as 300–450 HU, and the window center was set up to be 30–50 HU. Through a lung window, the morphological characters of burr tumor, lung interface and cavity of the patients’ lesion were evaluated. The window width was set to 1500–2000 HU, and the window center was set up as 450–550 HU. The diagnosed results were compared with pathological examination results of the patients or other results from laboratory inspection.

2.2 Research data and feature extraction

The pulmonary nodule images came from a comprehensive hospital. Fifty-six cases of lung cancer and 2240 pieces of CT images of pulmonary nodules with diameters $3 \leq d \leq 30$ mm were collected. The images were in DICOM format. The images were viewed through DICOM medical image browser software, and from the region of interest of lung CT images which includes the shape of pulmonary nodules, texture features and other features, etc. [10, 11]. The image analysis was executed on the key medical signs so as to achieve the feature extraction from the region of interest; the extracted features were: burr, lobulation, cavity, calcification, uniform density and sunken pleura. Details of the CT image features of solitary pulmonary nodule lung cancers studied are in Table 1 (among which: let male=1, female=0; lobulation=3, no lobulation=4; burr=5, no burr=6; cavity=7, no cavity=8; calcification=9, no calcification=10; uniform density=11, non-uniform density=12; pleural indentation=13; non-pleural indentation=14).

<table>
<thead>
<tr>
<th>Serial No</th>
<th>1</th>
<th>2</th>
<th>···</th>
<th>55</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender X1</td>
<td>0</td>
<td>0</td>
<td>···</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Age X2</td>
<td>52</td>
<td>63</td>
<td>···</td>
<td>54</td>
<td>73</td>
</tr>
<tr>
<td>Lobulation X3</td>
<td>3</td>
<td>3</td>
<td>···</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Burr X4</td>
<td>5</td>
<td>5</td>
<td>···</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Cavity X5</td>
<td>7</td>
<td>8</td>
<td>···</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Calcification X6</td>
<td>10</td>
<td>9</td>
<td>···</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Uniform density X7</td>
<td>12</td>
<td>11</td>
<td>···</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Sunken pleura X8</td>
<td>13</td>
<td>13</td>
<td>···</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

2.3 Application process of stepwise clustering

Gradual clustering works to make sum of squares of deviations within the sample groups reach a minimum standard. Through repeated adjustment of the number of individuals in each sample group, the optimization object, which is the maximum homogeneity (or the minimum heterogeneity) and maximum heterogeneity (or the minimum homogeneity) in sample groups, can be achieved. In the process of gradual clustering, this method has a rough classification of the samples at first, which is called initial classification, then repeated and continuous modification is executed in accordance with an optimization principle until reasonable classification is achieved [12].

In accordance with different analysis objects, clustering is divided into Q type and R type. Q type clustering is used to make classification processing on the sample, and R type clustering is to make classification processing on the variable [13]. This paper uses gradual clustering analysis on Q type samples.

2.3.1 Data conversion

As the dimensions of each factor in the system may not be the same, comparison is difficult to achieve, thus when the association analysis is in progress, treatment usually is carried out to make it nondimensionalized, so as to remove the influence brought by the each index dimension. A “Standardization” method is adopted in this paper to process the data:
\[ x_{ik} = \frac{y_{ik} - \bar{y}_k}{S_k} \quad (i = 1, 2, \ldots, n, k = 1, 2, \ldots, p) \]  

In which: \( \bar{y}_k = \frac{1}{n} \sum_{i=1}^{n} y_{ik} \), \( S_k^2 = \frac{1}{n} \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2 \), \( n \) is the sample number, \( p \) is the observed variable number, and in this paper, \( n=56, p=8 \). The standardized data are in Table 2.

### Table 2: Standardized data

<table>
<thead>
<tr>
<th>X1</th>
<th>0.7387</th>
<th>0.7387</th>
<th>\cdots</th>
<th>0.7387</th>
<th>0.7387</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2</td>
<td>1.2208</td>
<td>0.4580</td>
<td>\cdots</td>
<td>0.7860</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.1705</td>
</tr>
<tr>
<td>X3</td>
<td>-</td>
<td>-0.4337</td>
<td>\cdots</td>
<td>2.2647</td>
<td>-</td>
</tr>
<tr>
<td>X4</td>
<td>-</td>
<td>-0.6266</td>
<td>\cdots</td>
<td>1.5670</td>
<td>0.4337</td>
</tr>
<tr>
<td></td>
<td>0.6266</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0.3433</td>
<td>0.3433</td>
<td>\cdots</td>
<td>0.3433</td>
<td>0.3433</td>
</tr>
<tr>
<td>X6</td>
<td>0.2749</td>
<td>0.2749</td>
<td>\cdots</td>
<td>-3.5732</td>
<td>0.2749</td>
</tr>
<tr>
<td>X7</td>
<td>0.6268</td>
<td>0.6268</td>
<td>\cdots</td>
<td>0.6268</td>
<td>0.6268</td>
</tr>
<tr>
<td>X8</td>
<td>-</td>
<td>-1.5670</td>
<td>\cdots</td>
<td>0.6268</td>
<td>0.6268</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.5670</td>
</tr>
</tbody>
</table>

### 2.3.2 Initial clustering by rounding through conversion method

The method of integer transformation is adopted in initial classification. For each sample \( X_{ij} \), let

\[
SUM(i) = \sum_{j=1}^{m} X_{ij}
\]

\( SUM(i) \) represents the index variable value sum of each sample (\( m \) is the index variable number). If all samples are going to be classified, the calculation shall be made to each sample:

\[
\left\lfloor \frac{(K - 1)(SUM(i) - MI)}{MA - MI} \right\rfloor + 1
\]

If the integer adjacents to such a number is \( k \), then the sample \( X_i \) shall be classified to \( k \) category \((1 \leq k \leq K)\). We can obtain from Table 2, in this case, \( MI = \min SUM(i) = -9.054, MA = \max SUM(i) = 7.5542, [\cdot] \) means rounding operation. As for the selection of the \( k \) value, based on the medical acknowledge and the repeated computerized debugging tests, it is appropriate to divide 56 samples into three groups in the initial stage, and the initial clustering shall be executed in DPS software system in accordance with formula (2) and (3).

### 2.3.3 Selection of condensation point

The condensation point is the point representing the centre of the class to be formed. The selection of the condensation point can greatly influence the classification results. A centroid method [14] is adopted in this paper. Firstly, the objects are artificially divided into several categories, then the gravitational center of each category is calculated so as to be the gravitational center of clustering. The mean value of samples in such a category is taken as the condensation point, with formula:

\[
g_j = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ij}
\]

Among which, \( g(j = 1, 2, \cdots, m) \) shall be the barycentric coordinate of category \( k(1 \leq k \leq K) \), \( n_k \) is the sample number of category \( k \). Therefore, in accordance with formula (4), the gravity center of initial classification is obtained, then the condensation point is obtained, and the initial classification's barycentric coordinate is in Table 3.

### Table 3: Initial classification

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0</td>
<td>0.6481</td>
<td>1</td>
</tr>
<tr>
<td>X2</td>
<td>0.4545</td>
<td>0.5342</td>
<td>0.8485</td>
</tr>
<tr>
<td>X3</td>
<td>0</td>
<td>0.1481</td>
<td>1</td>
</tr>
<tr>
<td>X4</td>
<td>0</td>
<td>0.2778</td>
<td>1</td>
</tr>
<tr>
<td>X5</td>
<td>1</td>
<td>0.8889</td>
<td>1</td>
</tr>
<tr>
<td>X6</td>
<td>0</td>
<td>0.9994</td>
<td>1</td>
</tr>
<tr>
<td>X7</td>
<td>0</td>
<td>0.7222</td>
<td>1</td>
</tr>
<tr>
<td>X8</td>
<td>0</td>
<td>0.7222</td>
<td>1</td>
</tr>
</tbody>
</table>

### 2.3.4 Cluster all samples based on the latest condensation point

The objective function \( S \) is defined as:

\[
S(i) = \sum_{i=1}^{m} n_i (\bar{x}_i - \bar{X}) (\bar{x}_i - \bar{X})
\]

Here, \( n_i \) is the number of items in sample group \( i \), \( \bar{x}_i \) is the mean value, and \( \bar{X} \) is the sum mean value of sample \( N, m \) represents numbers of the grouping of \( N \) samples. \( N = \sum_{i=1}^{m} n_i, and S \) is the distance between sample and category condensation point.
The distance from each \( x_i \) sample to each category condensation point is calculated, and the sample is classified into the category occupied by the nearest condensation point.

### 2.3.5 Modification clustering, making the clustering reasonable

After the initial classification takes shape, it needs to be modified, step by step. The different methods of dynamic clustering are distinguished mainly by different principles of modification and classification. There are two methods of modification and classification, the one-by-one method and the group-by-group method [15–17]. This paper adopts the group-by-group method. After the initial condensation points are selected, each sample is classified according to its nearest condensation point. Each condensation point constitutes a class by itself, with its nearest points belonging to that class. The class’s centre of gravity is then recalculated, with the new value replacing the previous condensation point. Additional samples are then classified and the procedure is repeated until all samples have been placed into a class. When the calculated center of gravity is the same as the original condensation point, the process is stopped. If the center of gravity does not match the original condensation point, the previous steps are repeated according to objective function \( S \) till agreement is reached.

### 3 Analysis of results

Gradual clustering is judged based on a minimum sum of squares of deviations in the samples group. Higher homogeneity in a group is realized by repeatedly adjusting the iterations. This paper used data from 56 patients of solitary pulmonary nodule lung cancer as samples; their sex, age and CT image features are used as indices. The sample data were sorted into three classes by gradual clustering, as shown in Figure 1, the numbers on y-axis represent the individual samples.

Different features relating to the three classes of patient CT images are as follows:
1. Mostly female patients with ages between 50 and 65 years. CT images of solitary pulmonary nodule lung cancer for this group show that in the pulmonary nodules there are complete lobulation and burr, texture density is homogeneous, cavitation, calcification and pleural indentation are not found.
2. Mostly male patients with ages between 50 and 80 years. This group’s CT images show that in the pulmonary nodules there are complete lobulation and burr, texture density is homogeneous, cavitation, calcification and pleural indentation are not found.
3. Mostly male patients with ages between 50 and 80 years. This group’s CT images show that in the pulmonary nodules there are no lobulation and burr, texture density is homogeneous, cavitation, calcification and pleural indentation are not be found.

### 4 Discussion

It is not necessary for gradual clustering to calculate the similarity coefficient matrix between all samples. It is only necessary to calculate the distance between each sample and the center of clustering, which is the same as calcu-
lating the sum of squares of deviations. This process can therefore greatly shorten the calculation time and memory requirements of the computer so as to improve work efficiency. Our results show that with the gradual clustering analysis, CT images for patients with solitary pulmonary nodule lung cancer can be classified into three types according to the similarity of features of CT images. The method of gradual clustering is available for discovering the features of patients CT images similarities, distinguishing images features differences, satisfying the completeness, but not losing the information. All are beneficial to increase the accuracy of inspection and diagnosis, decrease the false positive rate of pulmonary nodule identification, and provide early information about pathological changes for doctors to help them understand medical features of CT images of solitary pulmonary nodule lung cancer.

Some errors may have arisen in our analysis; possible sources of error are discussed below.

1. A limited range of samples have been used. The gradual clustering analysis in this paper was restricted to lung cancer patients hospitalized in the comprehensive hospital. Additional samples from other areas, and additional CT image features would increase the significance of the results.

2. The outlier value and improper clustering variables have little influence on the clustering results of gradual clustering. Improper initial clustering can be repeatedly adjusted. The process is, however limited because gradual clustering results are very sensitive to the initial clustering.

3. The performance of clustering algorithm is closely linked with data, and there is no single algorithm that works for all cases. At present, each clustering algorithm put forward by the researchers has its own advantages, disadvantages and specific range of application. For similar data sets, use of different clustering algorithms result in different results of division.

In conclusion, the author of this paper has classified medical features in CT images medical features of patients with solitary pulmonary nodule lung cancer. The method of gradual clustering is applied to analysis CT image features, the clustering of patients' CT image medical features of solitary pulmonary nodule lung cancer can be obtained, which will be an important reference to doctor's diagnosis. Further research is needed into obtaining strong association rules with high reliability in order to improve the accuracy of disease diagnosis.

Acknowledgement: This project was supported by the National Spark Program funding project (No. 2015GA701023), Science and technology project of enriching the people of Ningbo city (No. 2015C10043, 2016A10041) and Zhejiang education program (No. 2015SCG087).

References


