Statistical methods to improve the accuracy of retrieving similar images in histopathological information database

ISHIBASHI, Yuichi
Graduate School of Environmental Science, Okayama University
3-1-1 Tsushima-naka Okayama, Japan, E-mail: ishibashi@ems.okayama-u.ac.jp
HARA, Atsuko
Kitasato University School of Medicine
Kitasato, Sagamihara, Kanagawa, 228-8555, JAPAN, E-mail: mlc52923@nifty.com
OKAYASU, Isao
Kitasato University School of Medicine
Kitasato, Sagamihara, Kanagawa, 228-8555, JAPAN, E-mail: isaokaya@med.kitasato-u.ac.jp
KURIHARA, Koji
Graduate School of Environmental Science, Okayama University
3-1-1 Tsushima-naka Okayama, Japan, E-mail: kurihara@ems.okayama-u.ac.jp

We have developed a histopathological information database system that enables retrieval of similar images. At present, the database stores the images of histopathological specimen including medical certificate text information, that are related to breast disease. The system retrieves cases which resemble the target image and calculates the probabilities for possible diseases connected to the target.

Digitization of images is necessary for image retrieval, therefore large specimen images are divided into many small images and Wavelet transformation is performed upon each small image. The small images that characterize the diseases are taken as training data, and the divided small images are identified by pattern recognition by the Neural Network. The result of this identification is used as a feature vector for a specimen image. Similar images are retrieved by comparing the feature vectors of the targeted image and the specimen images in the database.

A medical certificate describes the histopathological diagnostic process which is completed by pathological doctors. This diagnostic information is digitally transformed by the extraction of keywords and creates a dictionary by applying a text mining technique. The types of keywords that are involved in differential diagnosis of diseases are analyzed by canonical discriminant analysis, using the disease as a group and the keyword as a variable. This result enables the calculation of probabilities for possible diseases by Bayes' theorem including the keywords. Furthermore, the probabilities for possible diseases of the targeted image can be calculated by combining training data and keywords that represent training data.

In order to improve the accuracy of similar image retrieval, it is necessary to select appropriate training data. Characteristic images are selected as training data referencing the keywords that associate with the diseases mentioned in the above analysis. Interestingly, the principal component analysis detected a unique group of training data that resembled the other types or a distant data of the same category. Consolidation of training data and the addition of the deficient data with specific information, assisted the decrease in the misclassification of pattern recognition.

1. The method for the calculation of image feature vectors and image retrieval

The feature vectors are calculated through the digitization of images in order to retrieve images. A histopathological image of diagnostic size may contain not only the cancerous characteristic area, but also inflammatory or normal sections, such as interstitium, adipose, normal cells and so forth. Therefore, the discernable small images (128 x 128 pixels) are extracted and digitized by Wavelet transformation and used as training data. In Wavelet transformation an original small image is divided into 18 sub-bands at level 6. A

feature vector for a small image contains the variances of each sub-band and 2 mean values of color-difference signal of the image. $\mathbf{x} \in R^p (p = 20)$ is a feature vector, $y \in \{1,2,..,G\}$ is a label, and $\{(\mathbf{x}_1^0, y_1),...,(\mathbf{x}_m^0, y_m)\}$ is a set of m input data as training data and label. Presently, the database stores 17 types of breast diseases, this includes cancers and inflammations. Forty-seven types of training data $(G = 47)$ were extracted, due to different types of plural data existing in disease or normal tissue.

One hundred and fifty small images (128x128 pixels) are redundantly extracted from a large diagnostic image (1024 x704 pixels). Each small image is recognized by the neural network using training data, where the feature vector $\mathbf{C} = (c_1, c_2, ..., c_G)$ is counted upon for each training data c_i. As the accuracy of recognition in hierarchical neural network with many variables is of low quality, the LVQ (Learning Vector Quantization) was adopted due to its flexibility and the advantage it has with using a plethora of variables. Large images have already been diagnosed, producing a set of feature vectors and diagnostic results that are stored in the database.

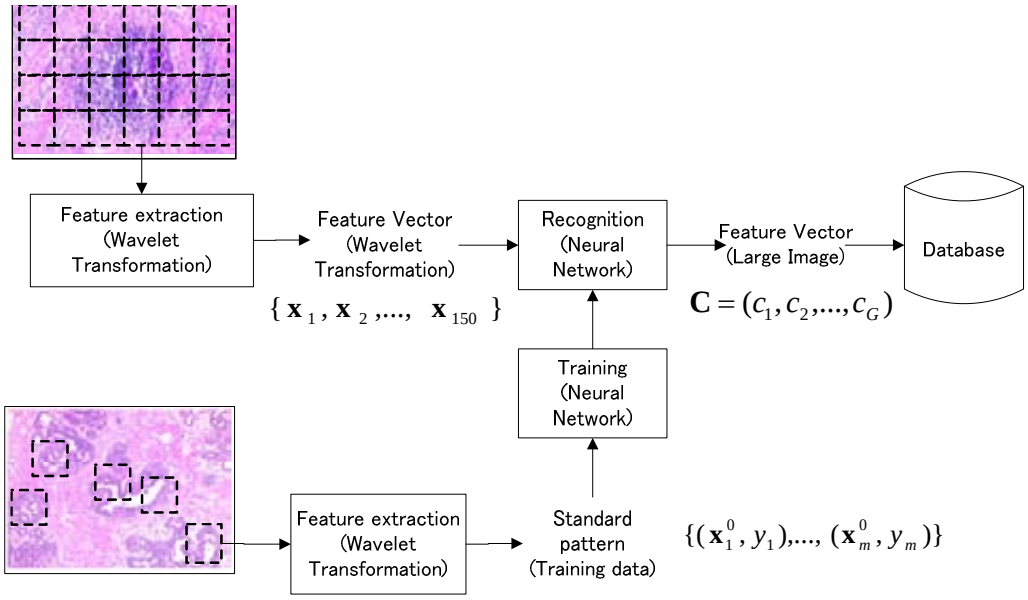

Figure 1 Digitization of images

When an undiagnosed pathological image (1024x704) is given, the feature vector of an entire image is calculated as in Figure 1. Similar image retrieval is performed as the feature vectors of the target image and the existing images in the database are classified by cluster analysis. Some classification methods were attempted, such as the method in which the short distance between two images is similar: hierarchical cluster analysis and non-hierarchical cluster analysis. The result of the classification in non-hierarchical k-means cluster analysis was the most suitable. Using these methods to specify the sample size of one cluster in non-hierarchical cluster analysis proved to be difficult, however the following algorithm makes the number of similar images be near N_0. In addition K_0 is (no. of existing images/2).

For i=K_0 to 1 step -1
  Result of classification = k-means clustering function (No. of clusters = i)
  if No. of similar images in a cluster in Result of classification >= N_0-1 then
    exit For
  end if
End For

2. Calculation of diagnosis probability by text mining

Various kinds of information are included in the pathological diagnosis report. These can range from a patients' age, materials, clinical diagnosis and pathological diagnosis names, acquisition method and

pathological observation. Please note that only the pathological observation and diagnosis name were adopted for text mining. A pathological observation describes the process of diagnosis by the pathological doctor. A dictionary is created by extracting keywords for diagnosis after the morphological analysis of original pathological diagnosis reports, and the numerical transformation of texts were performed. Canonical discriminant analysis with disease as a group and keyword as a variable, was adopted to analyze which keywords significantly contributed to the discrimination of disease.

It was proposed that keywords with selected variables (keywords) would be selected for diagnostic probability. Let the cause be diseases $C_1, C_2, \dots, C_n$ and the effects be keywords $w_{1,}, w_{2,} \dots, w_m$, thus the formula (1) is induced from Bayes' theorem. This formula is defined as diagnostic probability.

$$P\left(C_i \middle| w_{1,}, w_{2,} \dots, w_m\right) = \frac{P(C_i) \cdot P(w_1, w_2, \cdots, w_m | C_i)}{\sum_j P\left(C_j\right) \cdot P\left(w_1, w_2, \cdots, w_m \middle| C_j\right)} = \frac{P(C_i) \cdot \prod_k P(w_k | C_i)}{\sum_j P\left(C_j\right) \cdot \prod_k P\left(w_k \middle| C_j\right)} \quad (1)$$

We assume that keywords $w_{1,}, w_{2,} \dots, w_m$ are independent as formula (2), and the probability for each disease $P(C_i)$ is defined as the ratio of each disease in the database.

$$P\left(w_i, w_j \middle| C_k\right) = P(w_i | C_k) \cdot P\left(w_j \middle| C_k\right) \quad (2)$$

Fifteen keywords; for instance DCIS, solidtubular, and so forth, were extracted. A priori probability of a keyword for each disease $P(w_i | C_k)$ is calculated using the pathological observation in the database. This means that a priori probability is the ratio of the number of keywords in the document to the number of documents. The following example of formula (1) is the diagnostic probability of each disease for the pathological observation which is diagnosed IB2a3.

Example document: Grossly, white to gray solid tumor is confirmed in CD-area of the breast, measuring approximately 25 x 21mm. Histologically, the tumor is composed of round to polygonal shaped neoplastic cell proliferation with enlarged atypical nuclei. Tubular structure or compact alveolar configuration are seen in a small area, however, scirrhous infiltration with marked stromal fibrosis is mainly revealed, suggesting this case is an invasive scirrhous carcinoma. (snip)

Probabilities: IB2a3:0.374, IB2b3:0.307, IB2b1:0.148

3. Histopathological information database

The database contains pathological observations and diagnostic results as text information. The pointers are linked with large images (1024x704) which are extracted from an image on praparat, feature vectors of large images as numerical information, and so forth. The methods of similar image retrieval methods are as follows,

(i)To retrieve using only images.

(ii)To narrow the results of (i) by diagnostic probability linked with feature vector and keywords.

In image retrieval, a target image is processed through Wavelet transformation and pattern recognition by LVQ where the feature vector is created. Once created this feature vector of the new image and the feature vectors in the database are classified by cluster analysis. Similar images are in the cluster which now contains the new image.

Thumbnail images are created and displayed as a list of results. In the calculation of a diagnostic probability formula (1) which correspond with the training data are adapted, and some probable diseases are displayed. The previously retrieved images are reduced by using diagnostic probability.

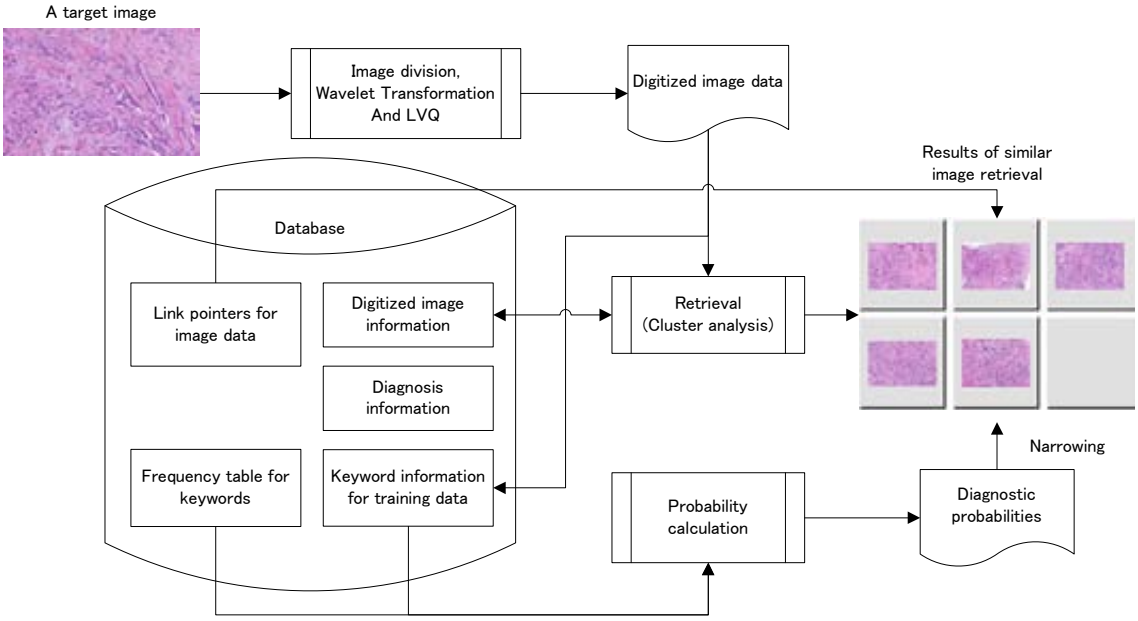

Figure 2 Structure of the database

4. Application of diagnostic probability in image retrieval

The training data which consist of characteristic sections of cancers or inflammations that represent configurations of papillary, medullary, scirrhous and so forth. Within the database, the training data is linked with these keywords. Keywords linked with training data i, $\{i \mid c_i \geq k_0, i = 1, \cdots G\}$ are extracted from the feature vector $\mathbf{C} = (c_1, c_2, ..., c_G)$, then the diagnostic probabilities for each disease are calculated using formula (1) in which the extracted keywords are substituted.

5. Improving of retrieval accuracy by reexamining of training data

To evaluate the accuracy of retrieval within the database, we assume the standard that the result of retrieval, contains the images with the same disease as the new image. Therefore, the ratio number of identical diseases and the number of retrieved images are used for evaluation. Based on this evaluation standard, improvement procedures of retrieval accuracy are mentioned as follows. It should be noted that these procedures are applicable to image retrieval only. This is due to the necessity of improving the accuracy of image retrieval before application of diagnostic probability.

5.1 Principal component analysis to evaluate training data

Assume a set of m Wavelet transformation values and labels $\{(\mathbf{x}_1^0, y_1), \cdots, (\mathbf{x}_m^0, y_m)\}$ is the training data, and the number of variables of $\mathbf{X} = [\mathbf{x}_1^0, \mathbf{x}_2^0, \cdots, \mathbf{x}_m^0]'$ is $p(=20)$. Therefore, the principal component analysis and is adopted to examine the relationship between training data in several dimensions. Also, we examine whether the same types neighborhood training data and the different types of training data that appear in the distance can be calculated by the principal component scores. Figure 3 shows the plot of principal component scores for 4 types of IB2a2 (Solid tubular carcinoma) training data. In this case, the cumulative contribution ratio has increased, making the second principal component 0.6026.

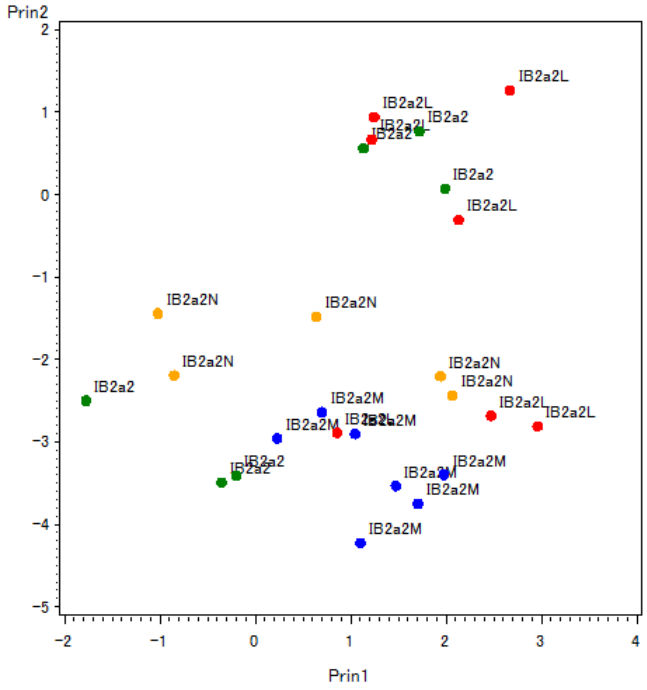

Figure 3 Plot of principal component scores.

5.2 Cross tabulation of training data and diseases

In the current database, 276 cases of image data are stored as the form in the feature vector $\mathbf{C} = (c_1, c_2, ..., c_G)$. Pattern recognition from the feature vector is counted and divided into small images for each training data. Considering that the feature vector for all cases is a 276 by G matrix; this matrix means a cross tabulation of relationship between training data and case data. Training data which contributes which is belonging to a disease were selected. Table 1 shows a section from a cross tabulation relating with IB2a2 and reveals the training data IB2a2L not contributing to disease IB2a2.

Table 1 Specific type of training data for IB2a2 (Part)

Case code	Kind of training data for IB2a2			
	IB2a2	IB2a2L	IB2a2M	IB2a2N
P06-01861-11C	47	0	0	0
P06-01217-16A	0	0	38	1
P06-01861-11B	34	0	0	0
P06-01861-11A	62	0	0	0
P06-01674-13G	0	0	105	11
P06-01674-13F	0	0	131	7
P06-01674-13D	0	0	99	7
P06-01674-13B	9	0	102	28
P06-01674-13A	0	0	101	24
P06-01217-16B	0	0	14	0
P06-01674-13E	0	0	97	17
P06-00954-19	0	0	128	8
P06-00507-10C	54	0	0	0
P06-01217-16C	0	0	32	0
P06-01674-13C	1	0	82	6

5.3 Evaluation of prediction error in neural network

Prediction and cross validation by a neural network were used for comparison between current and new training data sets. Table 2 shows the result of evaluation for 47 types of current training data.

Table 2 Accuracy of prediction (47 types of training data)

	Learning with all data	Cross validation
Training data grouped by disease	0.818	0.537
Individual training data types	0.779	0.430

5.4 Exchange of training data

Reducing the ambiguity is important for correct pattern recognition. Therefore, it is preferable that the same types of training data are in neighborhood and different types of training data separate. Based on principal component analysis and cross tabulation the result of selection, there were 30 types of training data by deletion and addition. Prediction and cross validation by neural network for 30 types was shown in table 3.

Table 3 Accuracy of prediction (30 types of training data)

	Learning with all data	Cross validation
Training data grouped by disease	0.873	0.580
Individual training data types	0.893	0.541

6 Results

Accuracy of similar image retrieval was measured using a new set of training data. This evaluation was carried out by the counts of the same disease in retrieved cases against the test case, which already had been diagnosed. Table 4 depicts the test data result of the retrieval of 27 test cases. The numbers of the same disease were identical, the number of different diseases were reduced, thus the rate of correct cases increased. The above method for the evaluation of training data proved sufficiently effective.

Table 4 Evaluation of retrieval accuracy

Kinds of training data	# of same diseases	# of retrieved cases	Ratio of correct cases
47(Before)	45	128	0.352
30(After)	45	116	0.388

REFERENCES

ARAI, K. (2000): Fundamental Theory on Wavelet Analysis, Morikita Shuppan (in Japanese).

ISHIBASHI, Y. et al. (2010): Statistical analysis of histopathological diagnosis reports with text mining, Joint meeting of Japan-Korea Special Conference of Statistics, 2010, Okayama, Japan. JIN, M. (2007): Data Science with R, Morikita Shuppan (in Japanese).

KANAMORI, T. et al. (2009): Pattern Recognition, Kyoritsu Shuppan (in Japanese).

KOHONEN, T. et al. (1995): LVQ PAK:The Learning Vector Quantization Program Package Technical Report, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, FINLAND.

KUMAGAI, J. and ITO, T. (2006):Histopathological diagnosis by image processing, Pathology and Clinic Vol.24, No.4. 2006 (in Japanese).

SAKAI, K. (2006): Introduction to the Image Processing and Pattern Recognition, Morikita Shuppan (in Japanese)..

TAMURA, H. (2002): Computer Image Processing, Ohmsha (in Japanese).