Journal of Computer and Communications, 2021, 9, 10-19
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Creating Knowledge-Based Diagnostic Models by Mining Textual Diagnostic Reports of SPECT Scans

Chuangui Cao[1,2], Chengcheng Han[1,2], Qiang Lin[1,2*]

[1]School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, China
[2]Key Laboratory of Streaming Data Computing Technologies and Application, Northwest Minzu University, Lanzhou, China
Email: 1689481324@qq.com, 2307115582@qq.com, *qiang.lin2010@hotmail.com

How to cite this paper: Cao, C.G., Han, C.C. and Lin, Q. (2021) Creating Knowledge-Based Diagnostic Models by Mining Textual Diagnostic Reports of SPECT Scans. *Journal of Computer and Communications*, **9**, 10-19.
https://doi.org/10.4236/jcc.2021.95002

Received: April 30, 2021
Accepted: May 11, 2021
Published: May 14, 2021

Abstract

Mining rich semantic information hidden in heterogeneous information network is one of the important tasks of data mining. Generally, a nuclear medicine text consists of the description of disease (*i.e.*, lesions) and diagnostic results. However, how to construct a computer-aided diagnostic model with a large number of medical texts is a challenging task. To automatically diagnose diseases with SPECT imaging, in this work, we create a knowledge-based diagnostic model by exploring the association between a disease and its properties. Firstly, an overview of nuclear medicine and data mining is presented. Second, the method of preprocessing textual nuclear medicine diagnostic reports is proposed. Last, the created diagnostic modes based on random forest and SVM are proposed. Experimental evaluation conducted real-world data of diagnostic reports of SPECT imaging demonstrates that our diagnostic models are workable and effective to automatically identify diseases with textual diagnostic reports.

Keywords

Text Classification, Nuclear Medicine, SPECT Imaging, Diagnostic Model, Random Forest, SVM

1. Introduction

With the continuous improvement of human living standards, the requirements for health conditions are getting higher and higher. In view of this, academia has also begun to pay attention to medical research. Bone metastasis of malignant tumor is the malignant tumor with the highest incidence of bone structure [1].

In the past decades, 99mTc-methylene bisphosphate (99mTc-MDP) whole-body planar bone imaging has been widely used in the diagnosis of bone metastases [2]. With the advancements of imaging technologies, medical imaging modalities have been transforming from structural imaging like Computed Tomography (CT) and Magnetic Resonance Imaging (MRT) to the current functional imaging including Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT). There also are hybrid imaging modalities that combine structural imaging and functional imaging together such as SPECT/CT and PET/MRI. Specifically, the hybrid imaging technology has been gradually applied to the diagnosis of bone metastases, significantly improving the accuracy of the diagnosis of bone metastases [3].

Different from the SPECT/CT imaging technology, the SPECT/CT report is a textual diagnosis report obtained from a patient after the SPECT/CT examination, which belongs to the class of nuclear medicine text data. The main content includes the doctors' description of the image and the diagnostic recommendations based on the doctors' experiential knowledge. Creating computer-aided diagnostic models play a key role in clinical applications of nuclear medicine with accelerated rich textural reports from various patients with different diseases.

In order to construct a nuclear medicine text diagnosis model, a text classification method based on traditional machine learning is proposed by leveraging the classical random forest and SVM models. The built models have been evaluated with a set of real-world data of SPECT nuclear medicine text, showing the workability and effectivity for identifying diseases.

The rest of this paper is organized as follows. In Section 2, we review related work on text mining and machine learning-based diagnostic models. In Section 3, we provide the used data and proposed method. And in Section 4, we conclude this work.

2. Related Work

As an important branch of data mining, text mining has done a lot of research up to now. In particular, medical text mining has important practical value. Among them, medical text mainly includes structured text and unstructured text. For the structured text, the main and fundamental work is to build a powerful semantic knowledge base [4]. A rule-based approach is often used for unstructured purposes [5]. It is mainly multi-domain oriented and has strong applicability, but the accuracy and completeness of the extraction results are generally low.

Li *et al.* [6] proposed a method that combines the information gain of the positions of words and the semantic computing based on HowNet to extract Chinese named entity relations, and presented a relation extraction method of Chinese named entities, called LSE, which is scalable, semi-supervised and domain independent. Jonnalagadda *et al.* [7] proposed a semantic partitioning method to advanced concept extraction in the medical field by using a discriminant classifier (CRF) to extract medical concepts from medical errors, clinical descriptions and trials. The Unified Medical Language System (UMLS) is used for semantic structuring and information extraction, and a method for automatically generating knowledge representation from natural language text data, SEREMED, is proposed [8].

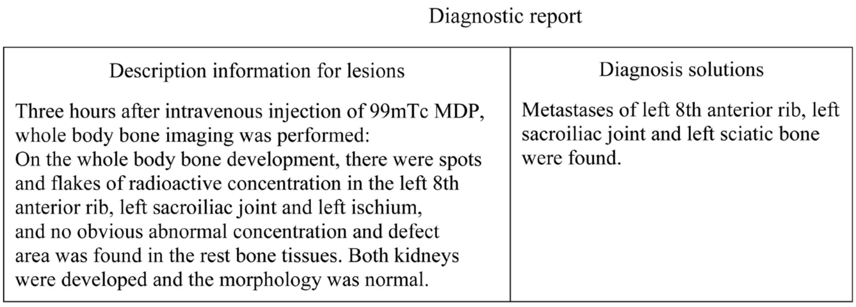
Shah *et al.* [9] proposed to develop an automated method for extracting the coded information from free text in electronic patient records. Onitilo *et al.* [10] developed an algorithm for identification of patients with type 2 diabetes and ascertainment of the date of diabetes onset for examination of the temporal relationship between diabetes and cancer using data in the electronic medical record (EMR). Yang *et al.* [11] used text clustering and keyword extraction to obtain the commonly used terminology in medical description language. As an important branch of data mining, text mining has done a lot of research so far. Filipe *et al.* [12] predicted future hospitalizations and discharges by analyzing the diagnostic data of patients in the early emergency department of a hospital. Kim and Delen [13] analyzed frequent keywords in top information journals and found that the Internet has changed the research methods in the field of medical research. Yang [14] studied a new ontology-based venous thromboembolic (VTE) risk assessment model, and verified the effectiveness of the model on real clinical data sets.

Existing work mentioned above focusing only on how to create a diagnostic model, while ignoring the characteristics of nuclear medicine text data itself. In other word, the unique lexical and grammatical characteristics of nuclear medicine text data was not considered by existing research efforts. In this work, we provide an in-depth analysis of diagnostic reports and develop reliable diagnostic models with textual data from SPECT imaging modalities.

3. Materials and Methods

3.1. Dataset and Preprocessing

The used diagnostic reports were collected in the process of diagnosing bone metastasis and related diseases in Department of Nuclear Medicine, Gansu Provincial Hospital from Jan. 2018 to Dec. 2019. For the examination of a patient, two SPECT images (*i.e.*, the anterior and posterior) and a corresponding diagnostic report in text format are recorded. **Figure 1** demonstrates an example of SPECT examination, with a male patient diagnosed with bone metastasis in ribs.



Diagnostic report

Description information for lesions	Diagnosis solutions
Three hours after intravenous injection of 99mTc MDP, whole body bone imaging was performed: On the whole body bone development, there were spots and flakes of radioactive concentration in the left 8th anterior rib, left sacroiliac joint and left ischium, and no obvious abnormal concentration and defect area was found in the rest bone tissues. Both kidneys were developed and the morphology was normal.	Metastases of left 8th anterior rib, left sacroiliac joint and left sciatic bone were found.

Figure 1. Illustration of diagnostic report from SPECT nuclear medicin examination.

As show in Figure 1, a diagnostic report text consists of two parts as follows:

- Descriptive information of lesions: The left column in Figure 1 shows the information of lesions in terms of position (joint) and status (normal) as well as the used radioactive tracer.
- Diagnostic solutions: The right column in Figure 1 provides the diagnostic solution done by nuclear medicine doctors.

The objective in this work is to develop a text classification method based on machine learning by exploring the diagnosis description and results. The used experimental data were collected from Department of Nuclear Medicine, Gansu Provincial Hospital, during 2017-2018. Table 1 outlines the used data of textual diagnostic reports in this work.

Given any SPECT examination cases, you can always find position, shape, level, state, and class information (or at least some of them). The first four are the description of the lesion itself, and the last is the description of the disease class of the lesion. First, select case description and diagnosis description from nuclear medicine text, which needs to screen out sensitive information such as name; secondly, clear the obvious wrong information in the text; thirdly, extract the features that can completely describe the diagnosis report from the text, including location, shape, level and state; Finally, the attributes are formally represented.

3.2. Diagnostic Models

3.2.1. Random Forest Based Model

We use decision tree as the base classifier of random forest. The decision tree generation process mainly uses information entropy theory. The Gini index is defined as Formula (1).

$$G\_i(D,a)=\sum_{v=1}^{V}\frac{\left|D^v\right|}{\left|D\right|}G\left(D^v\right) \tag{1}$$

where D^v refers to the sample set of the v-th class, and $G(D^v)$ is a random sample in the sample set.

The training set uses the model obtained from the test set to perform the final prediction classification. Figure 2 shows the working principle of the nuclear medicine text of the training set and the test set.

3.2.2. SVM Based Model

Different from the random forest model, SVM algorithm can solve the separation hyperplane with the largest geometric interval and the correct partition data set. As shown in Figure 3, the black circle and the white circle represent the correct text class and the wrong text class, respectively. The SVM looks for the

Table 1. The used data of textual diagnostic reports.

Diseases	Bone metastasis	Arthritis	Degenerative changes
Number	693	537	365

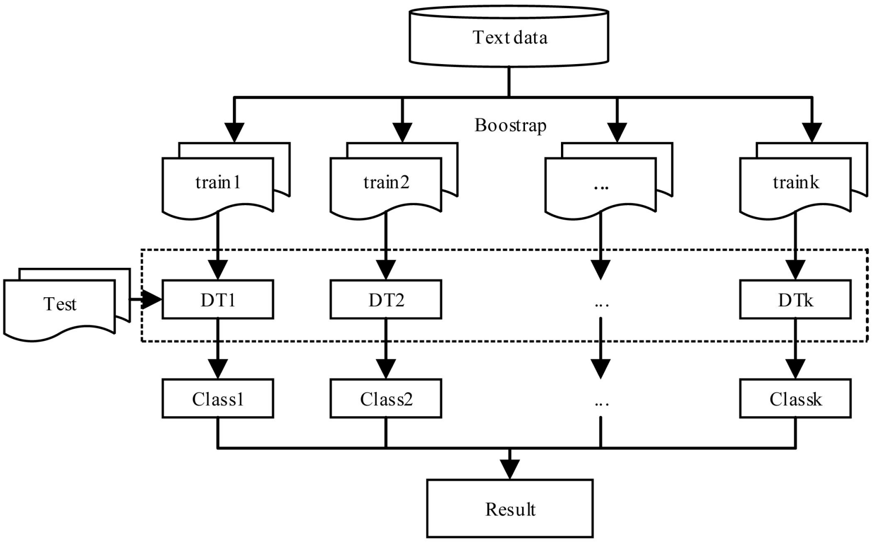


Figure 2. Preprocessing of random forest model decision.

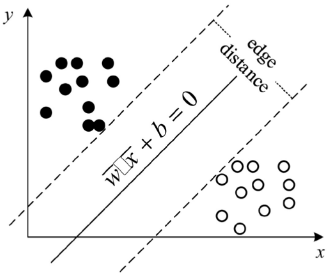


Figure 3. Support Vector Machines (SVM).

separated hyperplane that maximizes the margin between the dotted lines $\overline{w}\cdot\overline{x}+b=0$ ($\overline{w}$ is the plane normal vector, $\overline{x}$ is the feature quantity, and b is the displacement).

First, the representations obtained by extracting nuclear medicine text features form a 4-dimensional vector space and express it in vector form; then, train and learn the vector form training samples through the SVC method to obtain a classification model; finally, the classification model trains the test samples to obtain the classification results.

4. Experimental Evaluation

In this section, the SPECT diagnostic text obtained in nuclear medicine clinical diagnosis was applied to verify the validity of the experiment.

4.1. Experimental Setup

The evaluation index of this experiment selects OOB (Out of Bag) precision and accuracy. The training set is sampled with replacement (boostrap). In this process, 1/3 of the samples are not sampled each time. This 1/3 sample set is called OOB, which is used to estimate the generalization accuracy.

Different from OOB, accuracy, as another evaluation index in this experiment, refers to the closeness between the test result and the true value. The specific definition is as follows:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{2}$$

where TP and TN represent samples that are predicted correctly, the former is predicted positive samples, and the latter is predicted negative samples; similarly, FN and FP represent samples that are predicted incorrectly, the former is still predicted positive samples, and the latter is predicted negative samples.

4.2. Experimental Results

4.2.1. Random Forest Experiment Results

The number of trees in the forest (n_estimators) is an important parameter in the entire random forest, which affects the classification results. In order to clearly see the value range of the number of trees and discover the relationship between the number of trees and the OOB accuracy during the training process, Figure 4 shows the relationship between the number of trees and the OOB error rate.

The results show that the larger the number of trees at the beginning, the smaller the error; as the number of trees gradually increases to 20, the smaller the error fluctuation range, indicating that the greater the number of trees, the OOB error will not decrease.

In order to examine the maximum depth of the tree (max_depth) and the quality of the classification results, this experiment has set up three schemes (see Table 2).

The experimental results show that the algorithm classification result of Scheme 2# is the best one, and the accuracy of OOB is the highest. Table 3 shows the OOB accuracy obtained by the combination of n_estimators and max_depth. Finally, the evaluation index OOB of Scheme 2# is the highest.

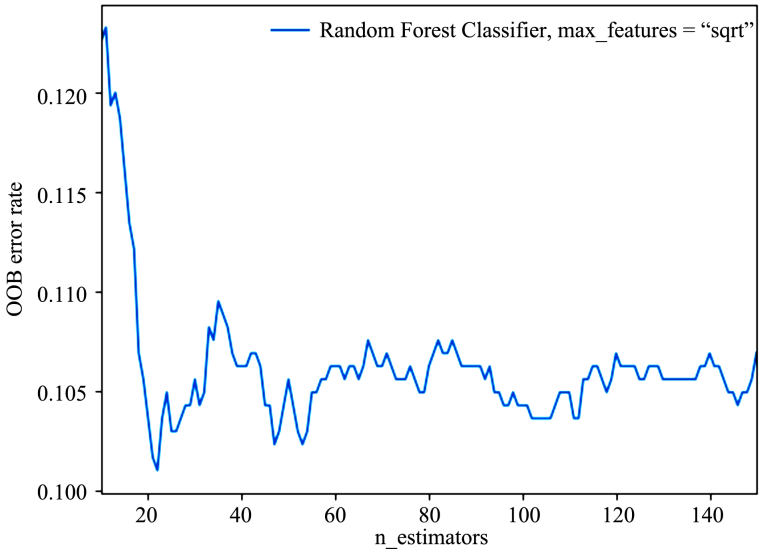


Figure 4. The relationship of n_estimators and OOB error rate.

Through the above adjustment process, the experimental results have obtained better results. From Figure 5 and Table 4 it can be seen that the OOB error rate is smaller than actual (about 0.005).

4.2.2. SVM Experiment Results

This part selects Support Vector Classification (SVC) in SVM to classify text, and introduces two kernel functions, *rbf* and *poly*. The classification results are shown in Table 5, where *cv* represents the number of iterations.

It can be seen from the table that when the rbf kernel function is selected and the number of iterations is 10, the classification effect of SVM is the best. But even so, compared with the classification results of random forest, it is far inferior to the classification effect of the latter. Therefore, this experiment finally chooses random forest to classify nuclear medicine texts.

Table 2. Various values for max_depth.

Scheme	max_depth
1#	6
2#	8
3#	10

Table 3. Major parameter value of random forest.

Scheme	Parameters	values	OOB	Acc
1	n_estimators max_depth	10 6	0.87	0.88
2	n_estimators max_depth	20 8	0.89	0.89

Table 4. Various values for max_depth.

Parameters	Value
n_estimators	20
max_depth	8
max_features	Auto
bootstrap	True
oob_score	True

Table 5. The results of Acc botained by SVM classification.

cv	rbf	poly
3	0.78	0.74
8	0.80	0.75
10	0.81	0.77
15	0.80	0.76

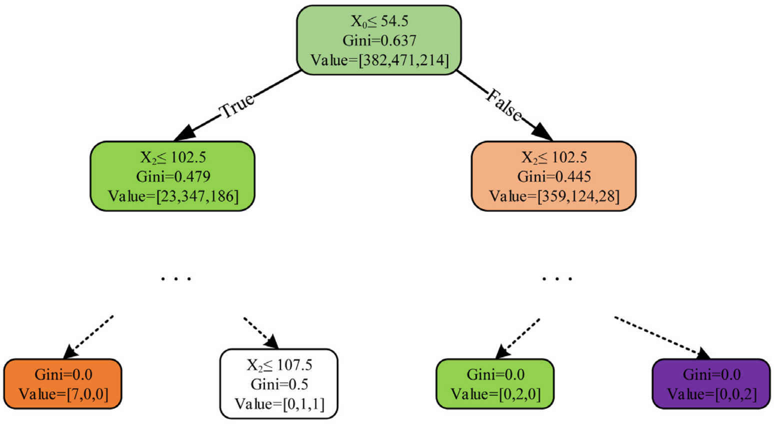


Figure 5. Illustration of classification by using decision tree based classifer.

4.3. Discussions

By adjusting the parameters of the random forest, a random forest model with classification effect can be obtained. Figure 5 shows a part of the result model of the decision tree, where orange represents arthritis, green represents bone metastasis, and purple represents degenerative changes. $X_{[0-3]}$ represents the four representations [position/shape/degree/state] in the quintuple, and formula 1 is used to calculate the information entropy for classification.

It can be seen from Figure 5 that the final classification result is white, and the content shows Value = [0, 1, 1], which does not belong to other diseases, which means that this node cannot continue to branch down, only as an unavoidable situation. This also explains why the OOB error rate in Table 3 is lower.

5. Conclusions

With the goal of building a nuclear medicine computer-aided diagnosis model, this paper has studied nuclear medicine text-assisted diagnosis model based on data mining, including:

First, the preprocessing process of nuclear medicine text is given, and the obvious errors or lack of speech in nuclear medicine text are dealt with. Secondly, a nuclear medicine text-assisted diagnosis model based on random forest has been proposed, especially considering the influence of the number of trees on the generalization accuracy. Finally, using real data from a hospital, the method proposed in this paper has been tested and verified, and the experimental results showed that higher evaluation results have been obtained.

In the future, we plan to extend our work in the following directions. First, we intend to collect more recordings of textual diagnostic reports to evaluate the proposed method. Second, we attempt to integrate domain knowledge to develop computer-aided diagnosis models.

Acknowledgements

This research was funded by Northwest Minzu University for Central University basic Scientific Research Operating Expenses Special Fund to Support the Graduate Program (Yxm2020100), the Natural Science Foundation of Gansu Province (20JR5RA511), the National Natural Science Foundation of China (61562075), the Fundamental Research Funds for the Central Universities (31920210013), and the Program for Innovative Research Team of SEAC ([2018] 98).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Vassiliou, V., Andreopoulos, D., Frangos, S., *et al.* (2011) Bone Metastases: Assessment of Therapeutic Response through Radiological and Nuclear Medicine Imaging Modalities. *Clinical Oncology* (*The Royal College of Radiologists*), **23**, 632-645. https://doi.org/10.1016/j.clon.2011.03.010

[2] Abikhzer, G., Gourevich, K., Kagna, O., *et al.* (2016) Whole-Body Bone SPECT in Breast Cancer Patients: The Future Bone Scan Protocol. *Nuclear Medicine Communications*, **37**, 247-253. https://doi.org/10.1097/MNM.0000000000000427

[3] Zhao, Z., Li, L., Li, F., *et al.* (2010) Single Photon Emission Computed Tomography/Spiral Computed Tomography Fusion Imaging for the Diagnosis of Bone Metastasis in Patients with Known Cancer. *Skeletal Radiology*, **39**, 147-153. https://doi.org/10.1007/s00256-009-0764-0

[4] Zhou, L. and Zhang, D. (2010) NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval. *Journal of the American Society for Information Science and Technology*, **54**, 115-123. https://doi.org/10.1002/asi.10193

[5] Chen, Y., Lee, S. and Huang, C.R. (2012) A Robust Web Personal Name Information Extraction System. *Expert Systems with Applications*, **39**, 2690-2699. https://doi.org/10.1016/j.eswa.2011.08.125

[6] Li, H.G., Wu, X.D., Li, Z., *et al.* (2013) A Relation Extraction Method of Chinese Named Entities Based on Location and Semantic Features. *Applied Intelligence*, 38, 1-15. https://doi.org/10.1007/s10489-012-0353-0

[7] Jonnalagadda, S., Cohen, T., Wu, S., *et al.* (2012) Enhancing Clinical Concept Extraction with Distributional Semantics. *Journal of Biomedical Informatics*, **45**, 129-140. https://doi.org/10.1016/j.jbi.2011.10.007

[8] Polkowski, C., Kaltenbach, B., Vogl, T. and Gruber-Rouh, T. (2008) Semantic Structuring of and Information Extraction from Medical Documents Using the UMLS. *Methods of Information in Medicine*, **47**, 425-434. https://doi.org/10.3414/ME0508

[9] Shah, A.D., Martinez, C. and Hemingway, H. (2012) The Freetext Matching Algorithm: A Computer Program to Extract Diagnoses and Causes of Death from Unstructured Text in Electronic Health Records. *BMC Medical Informatics & Decision Making*, **12**, Article No. 88. https://doi.org/10.1186/1472-6947-12-88

[10] Onitilo, A.A., Stankowski, R.V., Berg, R.L., *et al.* (2014) A Novel Method for Studying the Temporal Relationship between Type 2 Diabetes Mellitus and Cancer Using the Electronic Medical Record. *BMC Medical Informatics & Decision Making*, **14**, Article No. 38. https://doi.org/10.1186/1472-6947-14-38

[11] Yang, B., Nie, T.Z., Shen, D.R., *et al.* (2019) Approach of Structured Information Extraction for Medical Text Data. *Journal of Chinese Computer Systems*, **40**, 1479-1485.

[12] Lucini, F.R., Fogliatto, F.S., da Silveira, G.J.C., *et al.* (2017) Text Mining Approach to Predict Hospital Admissions Using Early Medical Records from the Emergency Department. *International Journal of Medical Informatics*, **100**, 1-8. https://doi.org/10.1016/j.ijmedinf.2017.01.001

[13] Kim, Y.M. and Delen, D. (2018) Medical Informatics Research Trend Analysis: A text Mining Approach. *Health Informatics Journal*, **24**, 432-452. https://doi.org/10.1177/1460458216678443

[14] Yang, Y., Wang, X., Huang, Y., *et al.* (2019) Ontology-Based Venous Thromboembolism Risk Assessment Model Developing from Medical Records. *BMC Medical Informatics and Decision Making*, **19**, Article No. 151. https://doi.org/10.1186/s12911-019-0856-2