Copyright © 2018 Tech Science Press CMC, vol.55, no.3, pp.419-433, 2018

Feature Selection Method Based on Class Discriminative Degree for Intelligent Medical Diagnosis

Shengqun Fang[1], Zhiping Cai[1, *], Wencheng Sun[1], Anfeng Liu[2], Fang Liu[3], Zhiyao Liang[4] and Guoyan Wang[5]

Abstract: By using efficient and timely medical diagnostic decision making, clinicians can positively impact the quality and cost of medical care. However, the high similarity of clinical manifestations between diseases and the limitation of clinicians' knowledge both bring much difficulty to decision making in diagnosis. Therefore, building a decision support system that can assist medical staff in diagnosing and treating diseases has lately received growing attentions in the medical domain. In this paper, we employ a multi-label classification framework to classify the Chinese electronic medical records to establish corresponding relation between the medical records and disease categories, and compare this method with the traditional medical expert system to verify the performance. To select the best subset of patient features, we propose a feature selection method based on the composition and distribution of symptoms in electronic medical records and compare it with the traditional feature selection methods such as chi-square test. We evaluate the feature selection methods and diagnostic models from two aspects, false negative rate (FNR) and accuracy. Extensive experiments have conducted on a real-world Chinese electronic medical record database. The evaluation results demonstrate that our proposed feature selection method can improve the accuracy and reduce the FNR compare to the traditional feature selection methods, and the multi-label classification framework have better accuracy and lower FNR than the traditional expert system.

Keywords: Medical expert system, EMR, multi-label classification, feature selection, class discriminative degree.

1 Introduction

Medical diagnostic decision making refers to the process of evaluating a patient complaint to develop a differential diagnosis, design a diagnostic evaluation, and arrive at a final diagnosis. The traditional method taken by doctors to diagnose the patients is mostly based on doctors' professional knowledge, work experience and test results.

[1] College of Computer, National University of Defense Technology, Changsha 410073, China.

[2] Computer Science and Technology, Central South University, Changsha 999078, China.

[3] School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China.

[4] Faculty of Information Technology, Macau University of Science and Technology, MACAU 410073, MO.

[5] Xuzhou University of Technology, Xuzhou 221002, China.

[*] Corresponding Author: Zhiping Cai. Email: zpcai@nudt.edu.cn.

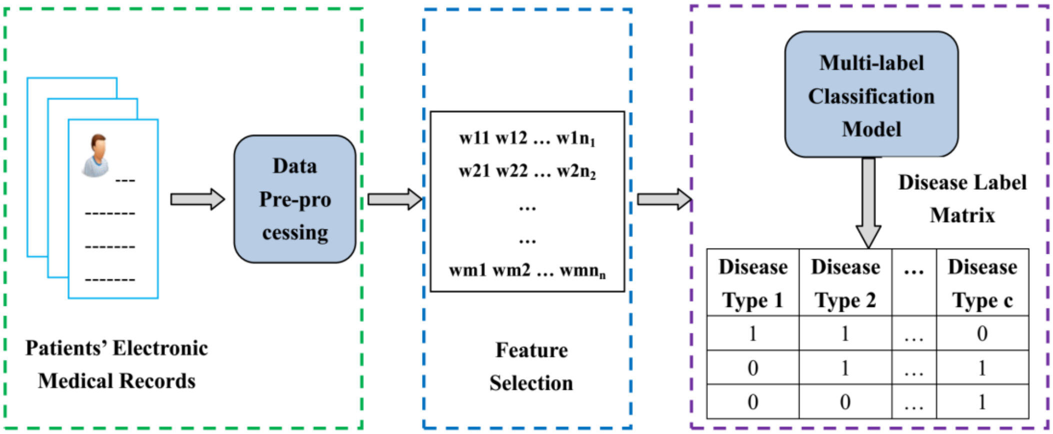
CMC. doi:10.3970/cmc.2018.02289 www.techscience.com/cmc

However, this diagnostic process is full of uncertainty and diagnostic results are closely related to the doctors' professional competence, especially for those complex medical records. At the same time, with the continuous development in the medical field, the disciplines of clinical medicine are getting more detailed. The doctors in different departments only pay attention to the subjects they have learned, which means that it is difficult for the doctors to conduct a comprehensive analysis on all aspects of the patients [Tang, Liu, Zhang et al. (2018)]. Accordingly, how to use information technology to develop a medical decision support system (MDSS) has become a hot research topic.

The first MDSS is an expert system based on knowledge base and rules of inference that utilizes computer technology to simulate the process of diagnosis and treatment that are usually done by medical experts. The United States started the earliest research on the medical expert system. The two most famous ones are the MYCIN expert system [Shortliffe, Davis, Axline et al. (1975)] developed by Stanford University in 1976 for the diagnosis and treatment of bacterial infections, and the Internist-I Internal Medicine Computer-Aided Diagnostic System [Miller, McNeil, Challinor et al. (1986)] developed by the University of Pittsburgh in 1982. With the development of information technology, the medical expert system becomes more complex. In Norouzi et al. [Norouzi, Yadollahpour, Mirbagheri et al. (2016)] build an adaptive neurofuzzy inference system using 10-year clinical records of newly diagnosed chronic kidney disease patients. In Ba¸ sçiftçi et al. [Ba¸ sçiftçi and Avuçlu (2018)] develop a new Medical Expert System (MES) which uses Reduced Rule Base to diagnose cancer risk according to the symptoms in an individual and a total of 13 symptoms. By controlling reduced rules, results are found more quickly. However, most of the expert systems are not intelligent enough. They only save the inference rules, professional knowledge and work experience summarized by the medical experts to the computer, and then get the diagnostic results through simple reasoning. On the one hand, it is not easy to put forward a sound and complete set of inference rules because the problems faced by doctors are very complicated. On the other hand, there is a limitation in the professional knowledge of the medical experts.

Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set [Liu, Dong, Liu et al. (2018)]. In the medical field, the documents refer to electronic medical records and the categories refer to disease types, therefore we can apply the data mining technologies on EMR [Sun, Cai, Liu et al. (2017)] to disease diagnosis. The workflow of traditional text categorization includes text preprocessing, text representation, feature selection and model construction, in which feature selection is extremely important primarily because it serves as a fundamental technique to direct the use of variables to what's most efficient and effective for a given classification model. However, features selected by traditional methods can be unrepresentative and highly sparse sometimes [Huang, Liu, Zhang et al. (2018)]. Under this circumstance, we need to acquire features manually. In recent years, the research idea of building language models using neural networks has become more and more mature. By using word embedding such as one-hot representation, combined with deep learning frameworks, feature selection can be omitted. In 2013, Mikolov et al. [Mikolov, Yih and Zweig (2013)] propose the word2vec model, which effectively solved 2289 the curse of dimensionality problem in the traditional one-hot word vector representation method and made further development of text classification using deep learning algorithms [Wang, Chang, Li et al. (2016); Johnson and Zhang (2015)].

In order to take advantage of machine learning techniques, this paper applies text classification algorithms to medical diagnosis, and proposes an intelligent diagnosis model. The demonstration of the proposed framework is shown in Fig. 1.



Disease Type 1	Disease Type 2	…	Disease Type c
1	1	…	0
0	1	…	1
0	0	…	1

Figure 1: Workflow demonstration of the proposed framework. The green box on the left contains the data pre-processing. The blue central box mainly select the features using chi-square test and the proposed method in this paper. The purple box on the right shows the process of building a model using a multi-label classification algorithm

The main contributions of this paper can be summarized as follows:

- Feature selection (also known as subset selection) [Chandrashekar and Sahin (2014); Sun, Cai, Li et al. (2018)] is a widely employed technique for reducing the dimensionality of the dataset. It aims to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results. We propose a feature selection method based on the composition and distribution of symptoms in electronic medical records and compare it with traditional feature selection methods such as chi-square test.
- Decision support for medical diagnosis is actually a multi-label classification problem, which means the patient, according to the medical records, is labeled as belonging to multiple disease classes. We adopt a variety of multi-label classification algorithms, such as binary relevance SVM (BR-SVM), multi-label kNN (MLkNN), etc. and compare the performance of different algorithms.
- To compare the performance of the proposed model with the medical expert system, we have invited medical experts from the cooperative hospital to construct a medical expert system.
- Extensive experiments have been conducted on a real-world Chinese medical record database obtained from the cooperative hospital. The experimental reports have shown that our proposed method is more effective for performing feature selection than the compared approaches and the intelligent diagnosis model is more effective

than the medical expert system in diagnosing diseases.

The rest of this paper is organized as follows: Related work will be reviewed in Section II. We will elaborate our method in detail in Section III, followed by evaluation reports in Section IV. We conclude the paper in Section V.

Table 1: Summarization of the dataset

	Disease name	Patients	Medical records
Training Set	Oral infection	248	8242
	Superficial infection	253	8526
	Urinary tract infection	245	9585
Testing Set	Oral infection	424	19074
	Superficial infection	410	19854
	Urinary tract infection	434	25147

2 Related work

2.1 Medical feature selection

There are no word boundaries in Chinese text. Therefore, unlike English, word segmentation is required before the feature selection for Chinese original text. We build a dictionary based on the names of Chinese medicines and diseases to improve the effect of word segmentation. There are three broad classes of feature selection algorithms [Chandrashekar and Sahin (2014)]: Wrapper, Filter and Embedded methods. Wrapper methods utilize the classifier as a black box to score the subsets of features based on their predictive power and the most commonly used classifier is SVM. Wrapper methods can be divided into deterministic and randomized [Saeys, Inza and Larrañaga (2007)]. The representative algorithms of the former include sequential forward selection (SFS) and sequential backward elimination (SBE), and the representative algorithms of the latter include Simulated annealing Randomized hill climbing and Genetic algorithms. Filter methods select features based on discriminating criteria that are relatively independent of classification and the commonly used criteria include simple correlation coefficients and mutual information. Filter methods can also be splitted into two categories [Saeys, Inza and Larrañaga (2007)]: univariate and multivariate. The commonly used feature selection methods, such as chi-square test, information gain and gain ratio belong to the former type and the latter type includes correlation-based feature selection (CFS) and Markov blanket filter (MBF). For Embedded methods, the feature selection algorithms are embedded as part of the learning algorithm and the typical algorithms include ID3, C4.5 and CART. In this paper, we summarize the characteristics of the Chinese electronic medical records and propose a feature selection method based on these characteristics.

2.2 Multi-label classification in disease diagnosis

Multi-label classification has been well studied in recent years [Zhao, Huang, Wang et al. (2015); Zhang and Wu (2015); Zhu, Li and Zhang (2016); Chen, Ye, Xing et al. (2017);

Tabatabaei, Dick and Xu (2017)]. The existing strategies could be roughly categorized into three families, based on the order of correlations that the learning techniques have considered [Zhang and Zhou (2014)]: (i) First-order strategy; (ii) Second-order strategy; (iii) High-order strategy. The latter two strategies consider the relevance between labels. In the medical field, patients may suffer from multiple diseases at the same time, so multi-label classification has attracted more and more research attention to this domain. Stefano et al. [Bromuri, Zufferey, Hennebert et al. (2014)] propose multi-label classification of multivariate time series contained in medical records of chronically ill patients. Many researches on multi-label classification have already pointed out that the potential correlations between labels have great impact on the classification performance. Wang et al. [Wang, Chang, Li et al. (2016)] propose an algorithm which can capture the disease relevance when labeling disease codes rather than making individual decision with respect to a specific disease, and the evaluation results demonstrate that the method improves multi-label classification results by successfully incorporating disease correlations. In this paper, we use the co-occurrence information of diseases as an indicator to measure the relevance of the diseases and set a threshold to determine whether the diseases are related or not. In our framework, we leverage BR-SVM and MLkNN to solve the multi-label classification problem. BR-SVM [Zhang, Cai, Liu et al. (2018)] attempts to convert the multi-label problem into single-label problem and MLkNN attempts to modify single-label classification algorithm to solve the multi-label classification directly.

Table 2: Co-occurrence frequencies of the 3 diseases

Disease A	**Disease B**	**Co-occurrence frequency**
Oral infection	Superficial infection	0.0015
Oral infection	Urinary tract infection	0.0118
Superficial infection	Urinary tract infection	0.0223

3 Methods

In this section, we first introduce the details of the data pre-processing method, followed by feature selection from the electronic medical records will be elaborated. A medical expert system and an intelligent diagnosis model based on the multi-label classification algorithms are subsequently built to solve the aforementioned problems.

3.1 Data pre-processing

Data pre-processing involves the following steps:

- Remove the negative phrases. There are many negative phrases in Chinese electronic medical records, for instance, “history of negative hepatitis”. However, these phrases have little or no effect on the diagnosis. Therefore we collect negative words frequently used in medical records and form a negative word list to eliminate these phrases.
- Word segmentation. Word segmentation is needed for electronic medical record before

feature selection. On the one hand, Chinese medical records often contain a lot of noises, for example, a large number of typos are recorded or different descriptions of the same symptom in different hospitals. On the other hand, there are a lot of medical terminologies, such as chronic sore throat. To improve the performance of word segmentation, we build a dictionary based on the names of Chinese medicines crawled from the official website of the State Food and Drug Administration and the names of Chinese diseases extracted from ICD-10 disease code.

Table 3: Symptoms and scores selected by the medical experts

	Symptoms	Score
Oral infection	oral film	100
	stomatitis	100
	abscess of mouth	100
Superficial infection	incisional infection	100
	the wound ached	20
	red and swollen	80
Oral infection	odynuria	40
	urinary frequency	30
	urinary tract infection	100

3.2 Feature selection

Through the analysis of Chinese electronic medical records, we find that these medical records have the following three characteristics.

- The characteristic words have low repetitions. Medical records do not emphasize semantic connotations by repeating the key words, so key symptoms and signs information do not appear many times.
- Electronic medical records of patients infected with similar diseases have high overlap degree of key symptom words. For example, the medical records of patients infected with urinary tract infection always contain dysuria, urinary frequency and urgency.
- The key symptom words between different diseases are exclusive. For example, the words such as wound infection and wound inflammation appear only in the medical records of patients infected with superficial infection.

According to these three characteristics, we calculate the representativeness of features for each disease and select the best subset of features.

The notations used in this method are first summarized to give a better understanding of the proposed method. A total of N types diseases, X_1, X_2, $\cdots$, X_N, respectively. R_i is the number of patients in X_i and W_i is the sum of occurrences of words in electronic medical records of X_i. For a word w in X_i, w_i is the occurrence number in medical records of X_i, and r_i is the number of patients infected with X_i containing the word w in the medical records, then the representativeness of w for X_i can be represented as:

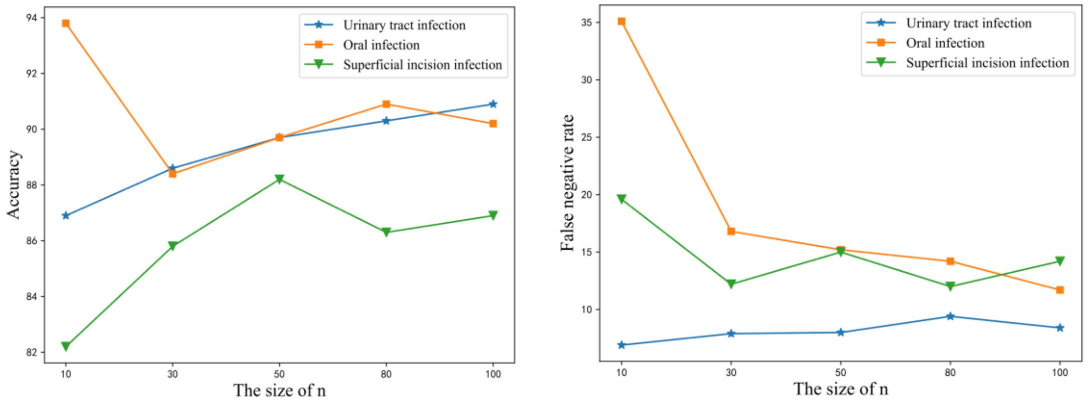
$$\mathrm{rep_w} = \log\left(\frac{W_i}{w_i}\right) * \left(\frac{r_i}{R_i}\right) * \left(\prod_{j=1, j\neq i}^{N} \log\frac{R_j}{r_j}\right) \tag{1}$$

where rep_w is the representativeness of w for X_i. For the first item log(W_i/w_i), this item will increase when w_i decreases which means the repetition of w is low. For the second item (r_i/R_i), this item will increase as r_i increases which means the number of patients with w in the medical records of X_i is high. As the part log(R_j/r_j) in third item, this part will decrease as r_j decreases which means the number of patients with w in the medical records of X_j is low. These three items in the formula correspond exactly to the three characteristics of the distribution and composition of the symptoms in the electronic medical record. After getting the representativeness of each word in X_i, we rank the representativeness and select top n words to form the feature vector for patients in X_i, where n is a parameter. We use the same method to select the symptoms of other diseases.

Table 4: Features selected by chi-square test and ours. Symptoms_A is selected by ours and Symptoms_B is selected by chi-square test

Diseases	Symptoms_A	Symptoms_B
Oral infection	tooth extraction	radiotherapy
	nasal douche	mucous coat
	oral hygiene	hyperemia
	gargle	sore throat
	oral film	gargle
Superficial infection	wound infection	wound
	wound inflammation	wound dressing
	incision infection	incision
	debridement and suturing	bind up
	purse-string suture	wound healing
Urinary tract infection	urinary tract infection	consciousness
	odynuria	catheter
	urine sediment	routine urine test
	urinary frequency	urine
	urinary urgency	urinary tract infection

The representativeness of a feature calculated by the traditional feature selection method is usually not used as the weight value of the feature. So after the feature selection, the weight value of each feature is calculated by the feature weighting method. The most commonly used feature weighting method is TF-IDF, where TF represents the term frequency and IDF represents the inverse document frequency. On the contrary, we directly use the representativeness calculated by the method proposed in this paper as the weight value of the features.



(a) Accuracy (b) FNR

Figure 2: These results are obtained using the BR-SVM. *n* represents the dimension of the feature vector. Fig. 2a shows the accuracy of different value of *n* and Fig. 2b shows the FNR of different value of *n*

3.3 Classification model construction

3.3.1 Medical expert system

The medical expert system collects the medical knowledge and the experience of medical experts from the cooperative hospital to extract symptoms for different diseases and score each symptom to build a database of disease symptoms. Then, according to the database, a diagnosis tree is constructed for each disease. For disease X_i, its diagnosis tree is T_i. The leaf nodes in the T_i are symptoms for X_i, and each branch node includes a score threshold. We first calculate the sum of the scores of the corresponding symptoms of leaf nodes under a branch node, then compare the sum with the threshold of this branch node. If the sum less than the threshold, then the score of the branch node is the sum, otherwise it is the threshold. Detailed steps to diagnose a patient using this expert system are as follows:

- Use ‘。’,‘！’ and‘？’ to cut each medical records into sentences, then use ‘，’ to cut each sentence into phrases. In the Chinese medical records, we think the words between two ‘，’ are the smallest unit of medical records.
- Delete phrases in which there are negative words belonging to the negative word list. If a phrase contains a negative word, the symptoms in the phrase should not be extracted.
- Extract the symptoms in the phrase and save these symptoms. We take out the symptoms in the database in turn and compare it with the content of the phrases.

- Create the diagnosis trees belong to the patient. We create the diagnosis tree of disease X_i for the patient based on the diagnosis tree T_i and the extracted symptoms belonging to X_i.
- Calculate the score of each disease according to the diagnosis trees for the patient, and compare the scores with a threshold. If the score of a disease is greater than the threshold, which means the patient is suffering from this disease. Note that if the sum of the score of the leaf nodes under a branch node is smaller than the threshold of this branch node, then the score of the branch node is the sum, otherwise it is the threshold. The threshold is set as 80 in this paper.

The expert system uses the knowledge and experience of medical experts to extract and score the symptoms of each disease, so the performance of the system is dependent on professional competence of the experts heavily. In order to reduce human intervention, this paper proposes a feature selection method based on the composition and distribution of symptoms in electronic medical records to extract symptoms for diseases automatically and combined with multi-label classification algorithm to construct a intelligent diagnosis system.

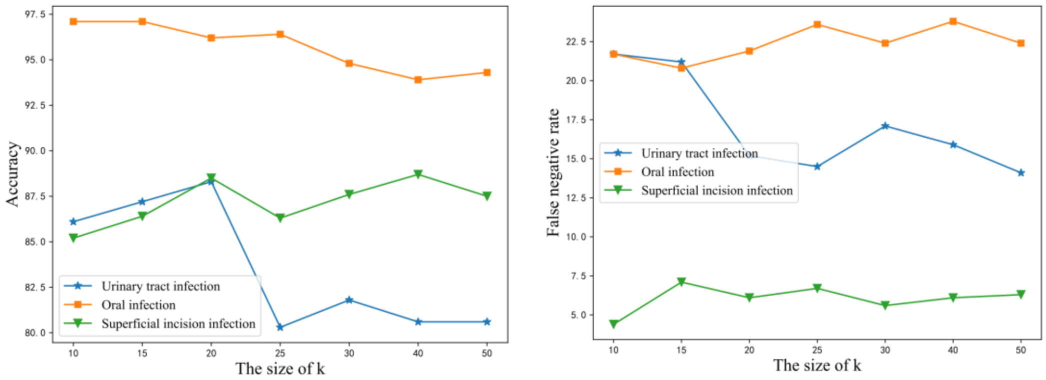
3.3.2 Intelligent diagnosis system based on multi-label classification

To determine whether the different diseases are related, we calculate the co-occurrence frequencies between diseases. Some notations are necessary to be explained before introducing the formula for calculating the co-occurrence frequency of two diseases. The class indicator matrix is represented as $D = [\mathbf{d}_1, \cdots, \mathbf{d}_m]_T \in R^{m*c}$. m is the number of training samples, c is number of diseases. $\mathbf{d}_i \in \{0,1\}^c$ is a c dimensional vector, if the i-th training sample belongs to the j-th disease, d_{ij} is 1, otherwise d_{ij} is 0, $i \in \{1, \cdots, m\}$ and $j \in \{1, \cdots, c\}$. The co-occurrence frequency for disease X_i and X_j is:

$$f_{ij} = \cos(p_i, p_j) = \frac{\langle p_i, p_j \rangle}{|p_i| * |p_j|} \tag{2}$$

where f_{ij} is the co-occurrence frequency of X_i and X_j. p_i is i-th column of D and $D = [p_1, \ldots, p_c]$, so p_i indicates the distribution of the i-th disease over the training data. We use cosine similarity to represent the relationships between X_i and X_j. The more correlated the X_i and X_j are, the higher the value of f_{ij} is.

Zufferey et al. [Zufferey, Hofer, Hennebert et al. (2015)] compare different multi-label classification algorithm for chronic disease classification and point out BR-SVM, which divides the multi-label classification problem into many binary classification problems, has achieved the best performance in terms of accuracy measured by Hamming loss. In this paper, we adopt two methods, BR-SVM and multi-label (MLkNN). As for BR-SVM, we use SVM algorithm to train a classifier for each of the three diseases. For a training set which is not linearly separable in original sample space, the SVM algorithm can use the kernel function to map the original samples to a higher dimensional feature space, and pervasively utilized kernel functions include [Bao, Wang and Qiu (2014)]: Linear kernel function (Linear), radial basis kernel function (RBF), polynomial kernel function (Polynomial) and the sigmoid kernel function. In this paper, we use basis kernel function according to the number of selected features and the number of samples. When training a classifier for a certain disease, patients infected with this disease are treated as positive samples and other patients are treated as negative samples, and the features used to train the classifier is the symptoms of this disease. Because the quantities of patients in each disease are close, this method will lead to the problem of imbalanced data. As over-sampling and under-sampling can solve this problem, we adopt these two strategies and compare the performance of them.



(a) Accuracy (b) FNR

Figure 3: These results are obtained using the MLkNN. *k* represents the number of nearest neighbors of samples. Fig. 3a shows the accuracy of different value of *k* and Fig. b shows the FNR of different value of *k*

Zhang et al. [Zhang and Zhou (2007)], they propose a multi-label lazy learning approach named MLkNN. For an unseen instance, its *k* nearest neighbors in the training set are firstly identified, after that, based on statistical information gained from the label sets of these neighboring instances, i.e. the number of neighboring instances belonging to each possible class, maximum a posteriori (MAP) principle is utilized to determine the label set for the unseen instance. The most important step of the framework is to find the *k* nearest neighbors of each instance in the training set. We use a similar pipeline in Zhang et al. [Zhang and Zhou (2007)] to diagnose each patient in testing set. For patient *P*, in order to calculate the distance between *P* and other patients in training set, we take a patient P_i out from the training set and convert the medical records of *P* and P_i into two n-dimensional vectors according to the symptoms corresponding to the disease Pi infected, and then calculate the cosine distance of the two vectors. After getting the cosine distance of patients in the training set and patient *P*, we can select the k nearest neighbors of *P*.

The training process of the intelligent diagnosis system is as follows:

- Get the training data. We obtain the patients' Chinese electronic medical records from the cooperative hospital and each patient suffers from only one disease.
- Initialize the word segmentation tool Ansj. Ansj is a commercial off-the-shelf tool and it is open source. We use this tool to import the custom dictionary and the stop word list.
- Use Ansj to convert medical records into a list of discrete words. First, we delete the

phrases containing negative words. We next utilize Ansj to segment the remaining phrases.

- Select features from the words. We use chi-square test method and the feature selection method proposed in this paper to select symptoms for different diseases.
- Construct the classification model. We use SVM and MLkNN to construct the intelligent diagnosis system.

Table 5: Performance comparison between our method and chi-square using MLkNN, over-sampling strategy and under-sampling strategy. Accuracy and FNR rate are used as metric

		MLkNN		Over-sampling strategy		Under-sampling strategy	
Criteria	**Diseases**	**Chi-square**	**Ours**	**Chi-square**	**Ours**	**Chi-square**	**Ours**
Accuracy	Oral infection	92.9%	96.2%	65.2%	89.2%	65.2%	89.4%
	Superficial infection	90.1%	88.5%	74.4%	87.9%	69.4%	85.7%
	Urinary tract infection	85.9%	88.3%	69.7%	89.4%	68.4%	88.3%
FNR	Oral infection	23.3%	21.9%	34.3%	13.5%	32.6%	16.2%
	Superficial infection	6.3%	6.1%	30.9%	13.0%	29.7%	14.2%
	Urinary tract infection	20.3%	15.2%	30.8%	9.6%	26.2%	6.3%

4 Experiments

There are already many public English electronic medical record datasets, but only Chinese electronic medical records are considered in this paper. We collect 398310 Chinese electronic medical records from 9602 patients from the cooperative hospital and divide these records into 27 types. We exclude 5 types according to the hospital standards and doctors' experience. The remaining types include admission records, chief physician rounds, first course records, ward-round records, surgery records, discharge records and so on.

Because the size of the dataset has a great effect on the classification result [Guo, Liu, Cai et al. (2018)], diseases with patient number less than 500 are ruled out. After disease filtering, we obtain three types of diseases. To train and test the multi-label classification algorithms, we split the patients in each type of disease into two parts, training set and testing set and the size of training set for each disease is close. Tab. 1 shows the data specifications. Note that a patient may suffer from multiple diseases at the same time, but the patients in the training set suffer from only one disease.

We use the co-occurrence frequencies to evaluate the correlation between the diseases in this paper. Tab. 2 shows the co-occurrence frequencies of the 3 diseases in Tab. 1. In Tab. 2, the co-occurrence frequency of any two diseases of the three diseases is very low, so we do not consider the correlation between the diseases when constructing the intelligent diagnosis system.

4.1 Evaluation criteria

To evaluate the performance, we adopt two criteria that are widely used in disease diagnosis: accuracy and FNR. For the disease X_i, the number of patients with this disease in the classification result is h_i, of which the number of patients predicted correctly is m_i.

- Accuracy: The accuracy for X_i refers to the ratio of mi to hi. From the definition, we can see that the larger the value of the accuracy is, the better the performance will be.
- FNR: The FNR for X_i refers to the ratio of (R_i–m_i) to R_i. From the definition, we can see that the smaller the FNR is, the better the performance will be.

4.2 Evaluation results

Tab. 3 lists some symptoms and scores of the 3 diseases selected by the medical experts from the cooperative hospital. Tab. 4 lists the top 5 most representative features for each disease selected by the chi-square test and the feature selection method proposed in this paper.

Table 6: Performance of the expert system

	Accuracy	**FNR**
Oral infection	84.6%	16.2%
Superficial infection	47.3%	4.88%
Urinary tract infection	53.8%	21.1%

Since there are two parameters, *n* and *k*, in our framework, we conduct two experiments to investigate performance variations with respect to different size of *n* and *k*. Performance variations with different size of *n* are depicted in Fig. 2. *n* varies in a range of {10, 30, 50, 80, 100}. Note that the bigger the accuracy value is and the smaller the FNR is, the better performance is. From the Fig. 2, we can observe that the best performance result is achieved when $n = 50$. As a result, we fix $n = 50$ in the rest of experiments. Performance variations with different size of *k* are depicted in Fig. 3. *k* varies in a range of {10, 15, 20, 25, 30, 40, 50}. From the Fig. 3, we can observe that the best performance result is identified when $k = 20$. As a result, we fix $k = 20$ in the rest of experiments. To consider the effectiveness of the feature selection method proposed in this paper, we compare it with the chi-square test method and report the results of the two types of features in Tab. 5. The representativeness of features calculated by the chi-square test is not used as the weight value of the features so we use the TF-IDF method to calculate the weight of the words selected by the chi-square test. From the table, we make the following observations: For the MLkNN, the accuracy and FNR of chi-square and our method are similar. But irrespective of the type of strategies used for SVM, our proposed method performs better than chi-square test. The accuracies of the three diseases are all close to 90% and FNR are all less than 20% obtained by our method. To verify the performance of the multi-label classification algorithm applied to the medical field, we compare it with the expert system. From the Tab. 5 and Tab. 6, we can observe that the intelligent diagnosis system based on multi-label classification performs much better than the expert system. The FNR of the expert system is close to the intelligent diagnosis system, but the accuracy is much worse.

5 Conclusion

The aim of this paper is to select the most representative features of different diseases, and then use these features to train classifiers to diagnose patients automatically. We obtain electronic medical records from the cooperative hospitals. The Chinese names of medicines are crawled from the official website of the State Food and Drug Administration and the Chinese names of diseases are extracted from ICD-10 disease code to build a dictionary. To solve the problem of data imbalance problem, we adopt the over-sampling and under-sampling strategies. With the goal of achieving acceptable accuracy and FNR for each disease, we propose a feature selection method based on intra class distribution and inter class distribution of symptoms. To verify the performance of the text classification algorithm applied to the medical field, we construct an expert system. Extensive experiments demonstrate that the proposed method selects feature of diseases more effectively than the traditional feature selection method and the performance of the intelligent diagnosis system based on multi-label classification performs much better than the expert system.

In future work, we would like to get the medical records of more types of diseases and consider the relationships between diseases when designing the feature selection method.

Acknowledgement: The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (No. 61379145) and the Joint Funds of CETC (Grant No. 20166141B08020101).

References

Bao, Y.; Wang, T.; Qiu, G. (2014): Research on applicability of svm kernel functions used in binary classification. *Proceedings of International Conference on Computer Science and Information Technology*, pp. 833-844.

Ba¸sçiftçi, F.; Avuçlu, E. (2018): An expert system design to diagnose cancer by using a new method reduced rule base. *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 113-120.

Bromuri, S.; Zufferey, D.; Hennebert, J.; Schumacher, M. (2014): Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. *Journal of Biomedical Informatics*, vol. 51, pp. 165-175.

Chandrashekar, G.; Sahin, F. (2014): A survey on feature selection methods.

Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28.

Chen, G.; Ye, D.; Xing, Z.; Chen, J.; Cambria, E. (2017): Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. *International Joint Conference on Neural Networks*, pp. 2377-2383.

Guo, Y.; Liu, F.; Cai, Z.; Xiao, N.; Zhao, Z. (2018): Edge-based efficient search over encrypted data mobile cloud storage. *Sensors*, vol. 18, no. 1, pp. 1189.

Johnson, R.; Zhang, T. (2015): Semi-supervised convolutional neural networks for text categorization via region embedding. *Advances in Neural Information Processing Systems*, vol. 28, pp. 919-927.

Liu, X.; Dong, M.; Liu, Y.; Liu, A.; Xiong, N. (2018): Construction low complexity and low delay CDS for big data codes dissemination. *Complexity*, vol. 2018.

Huang, M.; Liu, Y.; Zhang, N.; Xiong, N.; Liu, A. et al. (2017): A services routing based caching scheme for cloud assisted CRNs. *IEEE Access*, vol. 6, no. 1, pp. 15787-15805.

Milolov, T.; Yih, W. T.; Zweig, G. (2013): Linguistic regularities in continuous space word representations. Proceedings of NAACL-HLT, pp. 746-751.

Miller, R. A.; McNeil, M. A.; Challinor, S. M.; Masarie, F. E.; Myers, J. D. (1986): The internist-1/quick medical reference project-status report. *Western Journal of Medicine*, vol. 145, no. 6, pp. 816.

Norouzi, J.; Yadollahpour, A.; Mirbagheri, S. A.; Mazdeh, M. M.; Hosseini, S. A. (2016): Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. *Computational and Mathematical Methods in Medicine*, vol. 2016, no. 3, pp. 1-9.

Saeys, Y.; Inza, I.; Larrañaga, P. (2007): A review of feature selection techniques in bioinformatics. *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517.

Shortliffe, E. H.; Davis, R.; Axline, S. G.; Buchanan, B. G.; Green, C. C. et al. (1975): Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the mycin system. *Computers and Biomedical Research*, vol. 8, no. 4, pp. 303-320.

Sun, W.; Cai, Z.; Liu, F.; Fang, S.; Wang, G. (2017): A survey of data mining technology on electronic medical records. *2017 IEEE 19th International Conference on E-Health Networking, Applications and Services*, pp. 1-6.

Sun, W.; Cai, Z.; Li, Y.; Liu, F.; Fang, S. et al. (2018): Data processing and text mining technologies on electronic medical records: A review. *Journal of Healthcare Engineering*, vol. 2018.

Sun, W.; Cai, Z.; Li, Y.; Liu, F.; Fang, S. et al. (2018): Security and privacy in the medical Internet of Things. *Security and Communication Networks*, vol. 2018.

Tabatabaei, S. M.; Dick, S.; Xu, W. (2017): Toward non-intrusive load monitoring via multi-label classification. *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 26-40.

Tang, J.; Liu, A.; Zhang, J.; Xiong, N. N.; Zeng, Z. et al. (2018): A trust-based secure routing scheme using the traceback approach for energy-harvesting wireless sensor networks. *Sensors*, vol. 18, no. 3, pp. 751.

Wang, S.; Chang, X.; Li, X.; Long, G.; Yao, L. et al. (2016): Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3191-3202.

Zhang, H.; Cai, Z.; Liu, Q.; Xiao, Q.; Li, Y. et al. (2018): A survey on security-aware network measurement in SDN. *Security and Communication Networks*, vol. 2018.

Zhang, M.; Wu, L. (2015): Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107-120.

Zhang, M.; Zhou, Z. (2007): Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048.

Zhang, M.; Zhou, Z. (2014): A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819-1837.

Zhao, F.; Huang, Y.; Wang, L.; Tan, T. (2015): Deep semantic ranking based hashing for multi-label image retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1556-1564.

Zhu, X.; Li, X.; Zhang, S. (2016): Block-row sparse multiview multilabel learning for image classification. *IEEE Transactions on Cybernetics*, vol. 46, no. 2, pp. 450-461.

Zufferey, D.; Hofer, T.; Hennebert, J.; Schumacher, M.; Ingold, R. et al. (2015): Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Computers in Biology and Medicine*, vol. 65, pp. 34-43.