Mathematical Biosciences and Engineering

MBE, 18(2): 1000–1021.
DOI: 10.3934/mbe.2021054
Received: 29 September 2020
Accepted: 21 December 2020
Published: 05 January 2021

http://www.aimspress.com/journal/MBE

Research article

A method based on multi-standard active learning to recognize entities in electronic medical record

Qiao Pan*, Chen Huang and Dehua Chen

School of Computer Science and Technology, Donghua University, Shanghai 201620, China

* **Correspondence:** Email: panqiao@dhu.edu.cn; Tel: +862167792291.

Abstract: Deep neural networks（DNN）have achieved good results in the application of Named Entity Recognition (NER), but most of the DNN methods are based on large numbers of annotated data. Electronic Medical Record (EMR) belongs to text data of the specific professional field. The annotation of this kind of data needs experts with strong knowledge of the medical field and time labeling. To tackle the problems of professional medical areas, large data volume, and annotation difficulties of EMR, we propose a new method based on multi-standard active learning to recognize entities in EMR. Our approach uses three criteria: the number of labeled data, the cost of sentence annotation, and the balance of data sampling to determine the choice of active learning strategy. We put forward a more suitable way of uncertainty calculation and measurement rule of sentence annotation for NER's neural network model. Also, we use incremental training to speed up the iterative training in the process of active learning. Finally, the named entity experiment of breast clinical EMRs shows that it can achieve the same accuracy of NER results under the premise of obtaining the same sample's quality. Compared with the traditional supervised learning method of randomly selecting labeled data, the method proposed in this paper reduces the amount of data that needs to be labeled by 66.67%. Besides, an improved TF-IDF method based on Word2Vec is also proposed to vectorize the text by considering the word frequency.

Keywords: electronic medical records; multi-standard active learning; uncertainty; labeled costs; strategy choice

1. Introduction

Electronic medical record (EMR) records patients' critical clinical data during the examination

and treatment of diseases. It also contains a large amount of diagnosis and treatment information. By mining and using this information, EMR plays an essential role in developing the smart medical field [1–2]. Most EMRs are stored in the form of medical text entered in natural language. In this way, there are two questions to be considered: 1) It is difficult to directly obtain standardized and valuable data from these messy, redundant, and highly complex text data; 2) it is impossible to directly apply artificial-intelligence algorithms to further mine and analyze these data. Therefore, how to build the structure of EMR has become a hot issue in the era of artificial intelligence. Named entity recognition (NER) is the first step to design the structure of EMRs. However, NER's research in the medical field is still a big challenge due to the randomness, variability, and specialization of medical text data [3].

In recent years, deep neural networks (DNNs) have achieved good results in NER's application, but most deep learning methods are trained based on large numbers of annotated data [4]. Achieving excellent performance requires a large amount of labeled data for model training [4]. EMR belongs to the text data of the specific professional field. Thus, it requires experts with strong medical professional knowledge to annotate large amounts of training corpus, which costs a lot of workforce and time. In order to solve this problem, active learning has been proven as an effective tool for NER in clinical texts [5]. Controlling the process of labeling instances can reduce the workload of labeling medical records.

The core technology of active learning is strategy selection. Recent studies mainly focus on pool-based selection strategies, most of which use uncertainty selection strategies [6–8]. There are two uncertainty selection strategies: the one based on confidence and that based on information entropy [7]. The method based on confidence is simple to calculate, but it only considers the category with the highest posterior probability and ignores other categories' possibilities. So the effect is not significant in multi-class classification problems. In contrast, the approach based on information entropy considers the unlabeled samples' possibilities belonging to each category, so it is more suitable for multi-class classification problems. However, the computational complexity becomes higher because it requires plenty of logarithmic operations. Although the uncertainty selection strategy is the best strategy to reduce the labeling cost [9], literature [8] shows that when evaluating active learning and analyzing the algorithm, it should also consider the cost of manual labeling. The goal of active learning is to reduce the workforce of labeling data. The uncertainty-based method tries to solve this problem by minimizing the amount of required training data under the assumption that the cost of manual annotations is fixed. But in actual application scenarios, the workload of manual annotators is more complicated. For example, the cost varies depending on the entity type, the error type, the medical record sentence's length, and other factors. Therefore, the literature [8] designed a selection strategy considering the labor cost of labeling and applied it to the EMR. However, [8] does not fully consider the sparsity of medical short text data where the data distribution is unbalanced, which means it still has room for improvement in the entity recognition application from the medical text.

Several studies have recently shown that imbalanced data can easily undermine active learning performance [10–14]. In response to this problem, the literature [15–17] proposed solutions based on support vector machines to improve active learning efficiency to a certain extent. But these solutions take longer to run due to the high time complexity of support vector machine training. With the development of unsupervised learning, clustering technology brings hope to solve this problem [18]. Especially, the K-means clustering algorithm based on partition has become one of the most researched and applied clustering algorithms because of its simple, fast, and easy to expand

characteristics [18]. However, the traditional K-means clustering algorithm only considers the attribute characteristics of the sample, and it has certain blindness when ignoring the existence of priority information.

In this paper, we propose an active learning method based on a multi-standard combination strategy to solve the entity recognition task of EMR. The approach comprehensively considers three active-learning indicators: 1) Clustering is used to balance the sample data (i.e., data sparsity); 2) a new uncertainty selection strategy based on Gini impurity is proposed to reduce the amount of labeled data; 3) a combination strategy based on the relationship between uncertainty and annotation cost is designed by considering the practical application scenarios. Then by calculating the combined strategy score and selecting the instance with the higher score, the data labeling for training the entity recognition model is completed. The uncertainty calculation and measurement rule of sentence annotation for the NER's neural network model is proposed for entity recognition tasks. At present, most of the selection strategies of medical text data are single-standard active-learning methods [8,19], not based on multi-standard active learning that is widely used in image classification. As far as we know, we are the first to propose a medical text data selection strategy based on multi-standard active learning.

2. Related work

Our work involves the following research directions.

Data sparsity. For medical EMRs, an undeniable problem is the imbalance of data distribution [10–14]. In order to solve the impact of this problem on the active learning model, KSVM active learning algorithm [15], improved weighted SVM model [16], the active learning algorithm based on SVM hyperplane position correction [17] were proposed. These measures have improved the active learning efficiency to a certain extent. However, they take a long time to run because the time complexity of SVM training is high. Therefore, in literature [20], a clustering method is proposed to select samples to be annotated in the study of the active learning selection strategy. Specifically, the samples are first clustered into N categories, then the samples in each category are ranked through different methods, and the samples with the highest scores in each category are sorted to obtain the top M categories. The disadvantage of this method is that in some categories, there may be no samples selected, and this method may easily ignore the sparsity of sample distribution. In literature [21], a short text classification method based on clustering is also proposed, which includes K-means, singular value decomposition and affine propagation clustering algorithms, etc., among which K-means is the most frequently studied clustering algorithm. However, when solving the problem of imbalanced data distribution, the certain blindness of the traditional K-means clustering algorithm is that it only considers the attribute characteristics of the sample itself and ignores the existence of prior information.

Entity recognition. Since MUC (Message Understanding Conference) proposed the named entity recognition task [22], many methods have been proposed. Initially, entity recognition based on rules and dictionaries is a mainstream method [23–26]. However, it relies too much on manual dictionaries and regulations, which consumes many labor costs, and cannot adapt to new vocabulary emerging in the medical field. Then more methods based on machine learning are proposed [27–31], such as Bayesian Classification Model [31], Support Vector Machine (SVM) [32], Hidden Markov Model (HMM) [33], Maximum Entropy Markov Models (MEMM) [34], Conditional Random Fields

(CRF) [35] and many other models. But machine-learning methods still require researchers to extract effective features and formulate feature templates manually. In recent years, the rapid development of deep learning has attracted attention in entity recognition and the method has been widely used [36–38]. However, under typical training procedures, the advantages of deep learning will not be obvious when processing small data sets. Therefore, active learning has been increasingly used to reduce the amount of labeled data.

Table 1. Active learning for name entity recognition in electronic medical records.

Reference	Method	Selection criterion			Data
		Information and density	Sparsity and diversity	Cost	
Mahnoosh et al.[9]	Domain Knowledge Informatics (DKI).	√			i2b2/VA 2010[54], ShARe/CLEF 2013[55]
Mahnoosh et al.[51]	Least Confidence (LC), Information Density (IDen).	√			i2b2/VA2010,ShARe/CLEF 2013
Mahnoosh et al.[20]	Clustering And Representation Learning Sampling(CARLS).		√		i2b2/VA2010,ShARe/CLEF 2013
Wang et al.[52]	Combination of selection strategies based on Uncertainty (information Entropy) and Diversity.	√	√		Private data
Cheng et al.[53]	cost-sensitive active learning (CostAL).	√		√	Private data

Active learning. The core goal of active learning is to establish the criteria for selecting sample data that are most useful for the model [39–43]. Over the past few decades and so far, this problem has been the most concerned active-learning research point. Early researches included the member-based query method [21] and the stream-based sampling method [44]. The former ignores the actual distribution of the examples, and the latter's research lacks universality. Therefore, various pool-based sampling methods were subsequently proposed, including uncertainty-based sampling [45–47], version space-reduced sampling [48], and error-reduced sampling [49]. Although [50] can reduce version space, it may also select wild points in the data. Also, [49] has high time complexity, narrow application, and low-cost performance. Therefore, the most widely used method is based on uncertainty sampling. Similarly, the literature [9] shows a variety of selection strategies, and the experimental results also show that the performance of using uncertainty-based selection strategies to select examples with high information content is relatively the best. There are two major uncertainty-based selection strategies, which are respectively based on the least confidence(LC) and information entropy. The approach based on LC only considers the category with the highest

posterior probability and ignores other categories' possibilities. In contrast, the method based on information entropy considers the unlabeled samples' possibilities belonging to each class, so it is more applicable in multi-class problems. However, it requires many logarithmic operations, so that the computational complexity is very high.

In recent years, active learning selection strategies have been increasingly used for entity recognition in electronic medical records[9,20,51–53]. In [9], by comparing the performance of different AL query strategies for NER task in electronic medical records, they proposed a new active learning query strategy for information extraction, called Domain Knowledge Informativeness (DKI). Mahnoosh et al. [51] used the Least Confidence and Information Density as the active learning selection strategies while discusses the various evaluation metrics for active learning, taking into account the number of sentences, words and concepts. In Mahnoosh's another paper [20], they presented a novel active learning query strategy that takes into account sparsity and density using clustering. Given uncertainty and diversity, Wang et al. [52] proposed an active learning combination strategy for the NER in Chinese electronic medical records. However, one of the problems with these precious work is that they did not consider the time cost of labeling the instances. When there is a longer sentence in actual application scenarios, the information selected based on the uncertainty selection strategy requires a longer labeling time. Therefore, some active-learning methods based on cost sensitivity are studied to save the annotation cost [49,50,53]. These studies consider the cost of labeling data, but to a certain extent, the amount of information of the selected data is not necessarily important. We provide a table of comparison of previous works done in active learning and entity recognition for electronic medical records. The details are shown in Table 1.

3. Proposed approach

Different active learning strategies have different advantages in identifying which instance to query the given current classifier. This section proposes a method that combines different active-learning strategies' advantages in a balanced way. The proposed active learning method has four key components: an uncertainty measure, a cost-based measure, a data sparseness-based measure, and a combination strategy.

We denote U as the unlabeled dataset with n_u instances,

$$\mathrm{U} = \{\mathrm{x_j}\}_{\mathrm{j=1}}^{\mathrm{n_u}} \tag{1}$$

where x_j is the j-th instance. In this paper, we perform batch-mode active learning. At each iteration, a small batch of instances with size b will be selected from U to query their labels. Q is defined as a small batch of instances:

$$\mathrm{Q} = \{\mathrm{x_q}\}_{\mathrm{q=1}}^{\mathrm{b}} \tag{2}$$

In the following subsections, we will first introduce the batch-mode multi-standard active learning framework for NER of EMR, then propose the criterion for selecting instances and summarize the algorithm's main steps at last.

3.1. The framework

Our method focuses on batch-mode Multi-Standard Active Learning(MSAL). Figure 1 shows

the framework of the NER process based on MSAL. The way of MSAL is trained to recognize entities in EMR by iteratively selecting the training data and gradually improving the model performance to obtain strong generalization ability in a smaller data set. First, the unlabeled dataset is clustered and then the model is used to predict the clustered texts. According to the prediction results, uncertainty-based and cost-based selection strategies are used in each category to select the texts that meet the needs for annotation. The selected instance is then added to the labeled data set for the next training until the specified precision or the data amount is reached.

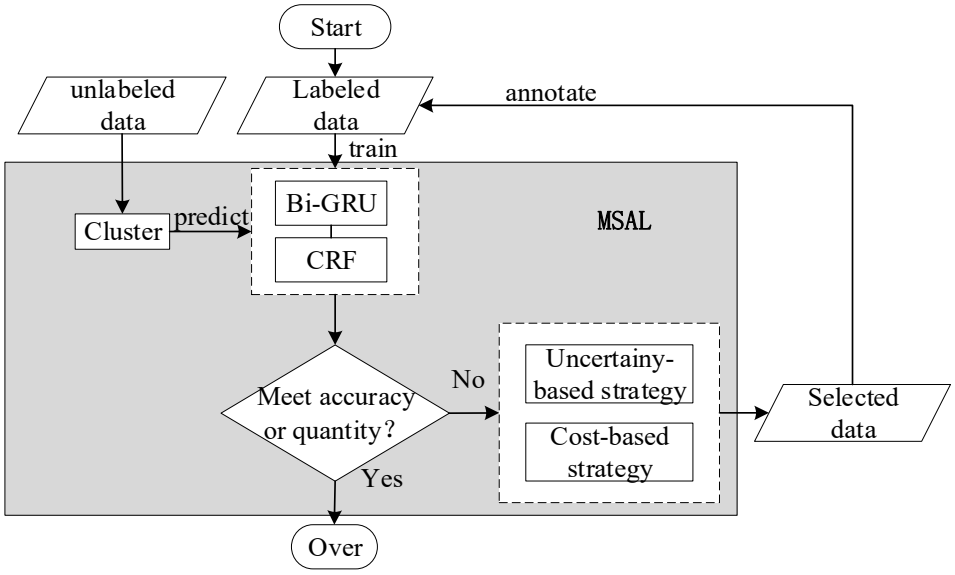

Figure 1. The framework of the Multi-Standard Active Learning(MSAL) model for named entity recognition.

3.2. The active selection criterion

The critical problem of improving the accuracy of entity recognition in medical EMR is to consider the reliability, the annotation cost of selecting instances, and the sparsity of medical short text data. On the one hand, the uncertainty measure is the best way to reduce the cost of annotation. Thus, adding a cost-based query strategy can further improve the active learning performance in EMR entity recognition. On the other hand, clustering can ensure data sampling balance for the sparsity of medical short text data. Therefore, we propose three criteria to estimate the usefulness of an instance on these two aspects. They are named clustering, uncertainty, and labeled cost, respectively. The workflow of criteria calculating is summarized in Figure 2.

3.2.1. Clustering

In order to ensure the instance distribution balance, it is necessary to evaluate the instances in the clustering category. In this paper, the improved TF-IDF [56] method based on Word2Vec [56] is used to vectorize the text, and then the k-means clustering algorithm is used to cluster the samples.

Although the word vectors obtained through Word2Vec retain the semantic information well, they may fail to express the importance of words to the text. While the TF-IDF algorithm only

considers the word frequency of text features in the corpus and ignores the context information [56]. Therefore, an improved vector calculation method is proposed to construct text vectors based on word vectors and weights:

$$\mathrm{vec(X)} = \sum_{x_i \in X} \mathrm{emb}(x_i) * (\mathrm{tf\_idf}) \tag{3}$$

where $emb(x_i)$ is an n-dimensional word vector generated by Word2Vec, * indicates dot product, tf is the frequency of the word in the sentence, idf is the frequency of reverse documents, the dimension of tf_idf is n * m, which represents the weight matrix of each word in each sentence. Therefore, $\mathrm{vec(X)}$ represents the importance of the non-sparse n-dimensional word vector in m text sentences.

Then the k-means algorithm is used to cluster the generated text vectors. First, K initial cluster centers are selected, then the Euclidean distance between each text vector and the cluster center is used as the similarity metric:

$$\mathrm{dis(vec(X_i),C_k)} = \sqrt{\sum_{p=1}^{n} \left(\mathrm{vec(X_i)_p - C_{kp}}\right)^2} \tag{4}$$

where X_i represents the i-th instance in the sample set, C_k represents the k-th cluster center, p is the word vector of the text vector, n is the word dimension of the text vector.

Accordingly, each sample is divided into the nearest cluster. After the division is completed, the new cluster center is recalculated using the mean value. The processes of dividing the samples and calculating the new cluster center execute cyclically until the clustering criterion function E converge:

$$\mathrm{E} = \sum_{i=1}^{k} \sum_{V \in \mathbb{R}^V} \|\mathrm{V} - \mathrm{C_k}\| \tag{5}$$

where V represents a text vector in all sample sets and C_k is one of the k clustering centers.

3.2.2. Uncertainty

The least confidence (LC), which is often used as the uncertainty measurement method, selects the instances with the most uncertain prediction results, i.e., the instances with the smallest maximum posterior probability. But LC only considers the category with the highest posterior probability and ignores the possibility of other categories. In information theory, information entropy is typically used to describe the uncertainty of information. Therefore, in strategy selection, entropy is also commonly used as a criterion to measure the uncertainty of an instance:

$$\mathrm{x^*} = \mathrm{argmax_{i=1,\ldots,n}} - \sum_i \mathrm{P(y_j|x_i)\,logP(y_j|x_i)} \tag{6}$$

where $P(y_j|x_i)$ represents the possibility that x_i belongs to the class j. The entropy-based method considers the possibility that an instance belongs to each class, so it is more applicable in multi-class situations. However, the entropy-based method involves a large number of logarithmic operations. When the amount of data is large, the calculation complexity is extremely high. Inspired by Gini impurity, which is like information entropy that can express the degree of data confusion, this paper uses Gini impurity to measure the instance's uncertainty. It can solve a large number of logarithmic operations and reduce the high computational complexity in information entropy.

Gini impurity refers to the expected error rate of the specific result in the collection. It can be

randomly applied to a data item, which is the simple probability that a random event becomes its opposite. Given a data set D, the impurity of it can be measured by the Gini index. The Gini index reflects the probability of inconsistency in the class labels of the two samples randomly selected from data set D. The smaller the probability of inconsistency is, the purer the set is. The uncertainty of one instance is defined as:

$$S_{Uncertainty} = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_k \, p_{k'} = 1 - \sum_{k=1}^{|K|} p_k^2 \tag{7}$$

where K represents the number of classes, p_k represents the probability that the instance belongs to the k-th class, and $p_{k'}$ represents the probability when the instance belongs to other classes except the k-th class. It can be seen from the above formula that the Gini index only needs to accumulate the probability that the instances belong to each class, which greatly reduces the amount of calculation. It is similar to entropy. When the instances are distributed in different classes with the same probability, the Gini index's value is the largest. When all instances belong to the same class, the Gini index is 0, which means the impurity is the lowest.

3.2.3. Labeled cost

According to [57], the concept of Annotation Rate (AR) is proposed for describing the actual annotation cost:

$$AR = \frac{\text{annotated samples used by AL to reach the target effectivenss}}{\text{annotated samples used by supervised learning}} \tag{8}$$

In the formula of AR, it is assumed that the annotation time of each sample is the same. However, in actual scenarios, each sample has a different annotation time which depends on the length of the sentence. Each expert has a certain reading speed. For a longer sentence, they need more time to annotate it. Therefore, when we describe the actual annotation cost of samples, we consider using the length of the sentence to represent it.

First, assuming there is reading cost, the average reading time for the selected sentence is Cr. Then, the prediction result of the selected sentence should be modified. If a particular value is labeled y_i, the probability of correcting the word can be approximate 1-P_{i,y_i}. Simultaneously, the time of the expert's revision is counted as the average word modification time Cw. Therefore, the labeling cost of each sentence can be defined as:

$$Cost = C_r * len(Sentence) + C_w * \sum_{i=1}^{n} (1 - P_{i,y_i}) \tag{9}$$

where p_{i,y_i} is the probability of the value label y_i with the i-th word obtained by softmax.

3.2.4. Combination strategy

As mentioned above, uncertainty is the best strategy to reduce the cost of annotation. But in the actual application scenario, it undoubtedly pays a high cost to obtain the most informative sample. We consider the selection strategy based on uncertainty and labeling cost to select the most suitable

model labeling instance. Therefore, the selected instance can be applied to the actual scene and improve classification performance.

It can be easily found that the active learning algorithm needs a longer labeling time to select more informative instances. If we want to acquire a higher accuracy rate, we need to set a higher weight ratio to the selection strategy based on uncertainty; if we're going to reduce the time for labeling instances, we have to lose the accuracy. When we use the selection strategy to select instances actively, we try our best to reduce the labeling time and ensure accuracy. Formally, we employ a trade-off parameter to balance the two criteria as the iterations progress:

$$\mathrm{score}(\mathrm{x}) = \mathrm{S}_{\mathrm{Uncertainly}}/\beta\mathrm{Cost} \tag{10}$$

where β is a trade-off parameter. Finally, we select the batch of instances with the highest score $(\cdot)$ value to query their labels and further optimize the classifier.

3.3. The Multi-Standard Active Learning (MSAL) algorithm

The main steps of the proposed MSAL algorithm are summarized in Algorithm 1.

Algorithm 1. The MSAL algorithm.

Algorithm 1: The MSAL algorithm

Input:
L:The labeled set
U: The unlabeled set with n_u instances
T: Train set
C: Cluster category
Q: The selected set from the unlabeled set with b instances

Output:
M: Entity recognition model

```
1: Repeat
2:     Train the model M with L
3:     Predict U with the model M
4:     for c in C:
5:         Select a batch of instances Q from U with largest score
6:         annotate the instances for Q
7:         T=random_b(L)+ Q
8:         L=L+Q
9:         U=U-Q
10: until query budget or expected performance reached
11: return selected model M
```

In the above algorithm, the initial labeled data set is selected by random extraction.

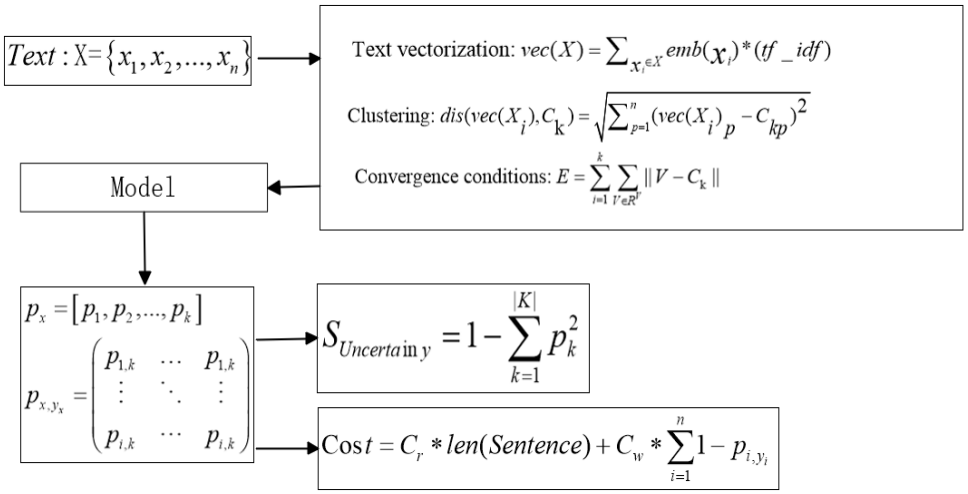

Figure 2. The criteria for active selection.

4. Experiments

In order to evaluate the entity recognition model based on active learning, the accuracy rate is selected as the evaluation standard of the experiment in this paper. Accuracy is widely used in information retrieval and statistical classification researches. In entity recognition, the calculation of this indicator is defined as follows:

$$\text{Accuracy} = \frac{\text{Intersection}}{\text{Entities extracted from the model}} \tag{11}$$

where "intersection" refers to the number of intersections between the entities extracted by the model and the entities actually in the data set, and "entities extracted by the model" refers to the number of entities extracted by the model.

4.1. Dataset

The experimental data used in this paper are 4000 copies of EMRs with breast disease from a top-three hospital in Shanghai, China. The data scale and the numbers of entities are shown in Table 2 and Table 3, respectively.

Table 2. The statistics of experimental data scale.

number of characters	number of short sentences	number of long sentences	number of paragraphs
7971556	606508	228477	32142

In this experiment, 3200 copies of these medical records were used as training data, and 800 copies were used as test data to analyze the model's accuracy. Active learning is a process of increasing the training data set. The training set's initial size was 35000 short sentences (about 300 medical records). It iterated by adding 17500 short sentences (approximately 150 medical records) per round.

4.2. Model selection

It is the most common way to treat the NER task of character-level Chinese EMRs as a sequence annotation problem. Several commonly used neural network architectures for sequence annotation tasks are shown in Table 4.

With continuous development of deep learning and NER, various improved structures have emerged, such as Bi-LSTM, Bi-GRU, Bi-RNN, etc., which learn contextual information from both forward and backward directions. CRF can automatically learn constraints from training data and improve the effectiveness of predicting labels. Therefore, it is more and more widely used in combination with models such as Bi-LSTM. Among them, the Bi-LSTM-CRF model is currently the most widely used NER model.

Although the LSTM networks is the most popular type of neural networks in Name Entity Recognition, GRU can also solve the problems of long-term memory and back propagation gradient and can also achieve similar performance as LSTM. Compared with LSTM, GRU is easier to converge, because its neural unit structure is simpler than that of LSTM. Therefore, it is more suitable for active learning, which requires repeated iteration, and can effectively improve the training speed. For this reason, the Bi-GRU-CRF model is used as the training model for NER in this paper.

Table 3. Numbers of experimental data entities.

Name of entities	Number of entities
Disease and diagnosis (DIS)	75,121
Test (TES)	22,059
Examine (EXA)	9589
Operation (OPE)	17,931
Medicine (MED)	36,034
Anatomic site (ANA)	154,972
Total number of entities	315,706

Table 4. Neural network architecture for sequence labeling tasks.

Name	Description	Reference
RNN	The disappearance of the gradient makes it impossible for long-time dependence.	Nguyen et al. (2016) [58]
LSTM	The "forgotten gate" is introduced, which effectively solves the problem of gradient disappearance. It is currently the most widely used neural network, but it has too many parameters, 4 times that of RNN, and there is a risk of over-fitting.	Huang et al. (2015) [59]
GRU	The few parameters can effectively reduce the risk of over-fitting and accelerate the speed of model convergence while achieving similar effects to LSTM.	Yang et al. (2016) [60]

4.3. Parameter Settings

The experiment determined the optimal parameters through trial and error. The hyperparameter settings used in the experiment are shown in Table 5.

4.4. Experimental results

In this part, the impacts of samples selected at different levels of sampling granularity on the training of entity recognition models are firstly studied. Then the uncertainty selection strategy based on Gini impurity proposed in this article and the uncertainty selection strategy which is most commonly used are compared. In order to evaluate the impact of different indicators on the performance of multi-standard active learning methods, ablation experiments are implemented.

Table 5. Experimental parameter settings.

name	value
Word vector dimension	200
batch size	3500
Number of GRU units	300
dropout	0.5
Learning rate	0.01
Number of iterations	50
Number of cluster categories	10

4.4.1. Comparative experiment of different sampling granularities

This experiment studies the accuracy of the entity recognition model based on the Gini impurity selection strategy under three sampling granularities of short sentence-level, long sentence-level and paragraph-level, so as to select the sampling granularity that can make the model with the highest accuracy. The baseline method of this experiment is a random extraction method, which is compared with three methods respectively based on short sentence-level Gini impurity, long sentence-level Gini impurity, and paragraph-level Gini impurity. The comparison of the entity recognition accuracy based on the selection strategy of Gini impurity under three sampling granularities is shown in Figure 3.

It can be seen from Figure 3 that the accuracy curve of the random sampling method is more tortuous. Because of the randomness of data selection, data that has a negative impact on the model may be added, which may compromise the accuracy of the model. All the other three levels have promoted the training of the model, but their effects are different. Although the sampling based on the paragraph level has a steady improvement effect on the accuracy of the model, the improvement rate is not as good as the other two. Sampling based on short sentences quickly improves the performance of the model at first, but it is not as good as sampling based on long sentences in the later stage. The division of short sentences is more detailed and can eliminate useless texts more accurately. However, compared with long sentences, short sentences may break up a complete sentence and lose part of the context, making the model lose a certain amount of information and leading to poor effects in the later stage. Therefore, the experiments in this article are based on long

sequences, but the training data is converted into short sentences before calculation.

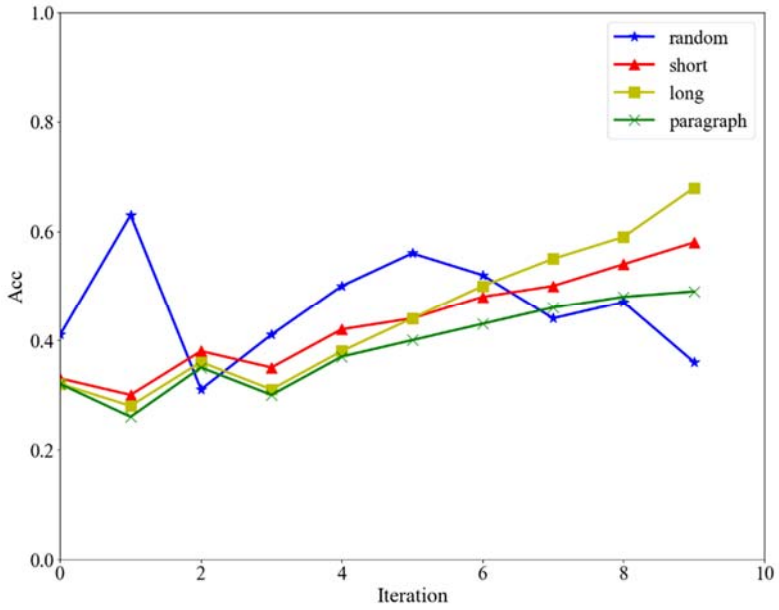

Figure 3. The accuracy of the Gini impurity based on different sampling granularity.

4.4.2. Study of the uncertainty

This experiment compares the uncertainty selection strategy based on information entropy (ENTROPY), confidence(LC), Gini impurity (GINI) proposed in this paper, and random method (RANDOM). The comparison of the accuracy is shown in Figure 4. The complexities of two uncertainty selection strategies (ENTROPY) and (GINI) are plotted in Figure 5.

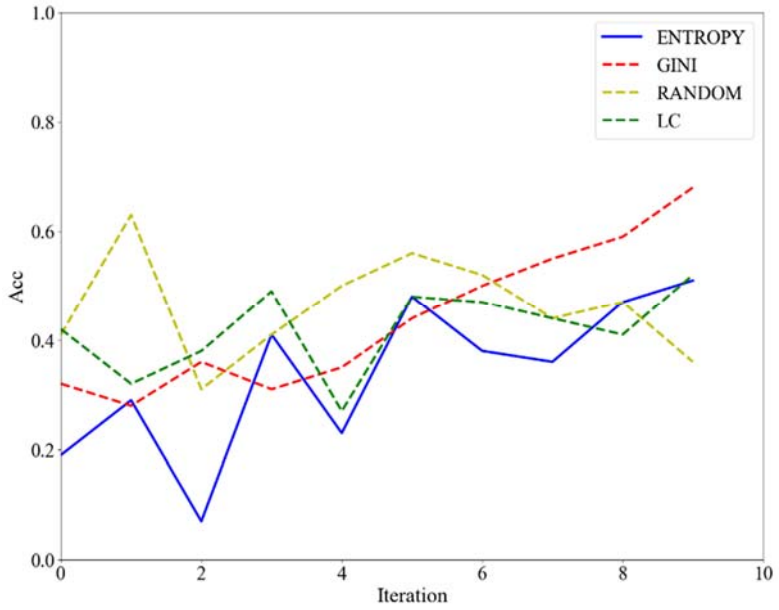

Figure 4. The accuracy of the selection strategies based on uncertainty.

It can be seen from Figure 4 that when the same amount of training data increases in iterations, the accuracy of the Gini impurity is higher than that of information entropy and confidence. Meanwhile, the random selection strategy's accuracy is not stable, decreasing with the increasing number of iterations.

Figure 5 shows the aspect of algorithm complexity. The computational complexity of the uncertainty selection strategy based on Gini impurity is much lower than that of information entropy. We can find that with the number of data increases, the gap between the two complexities will become more and more obvious.

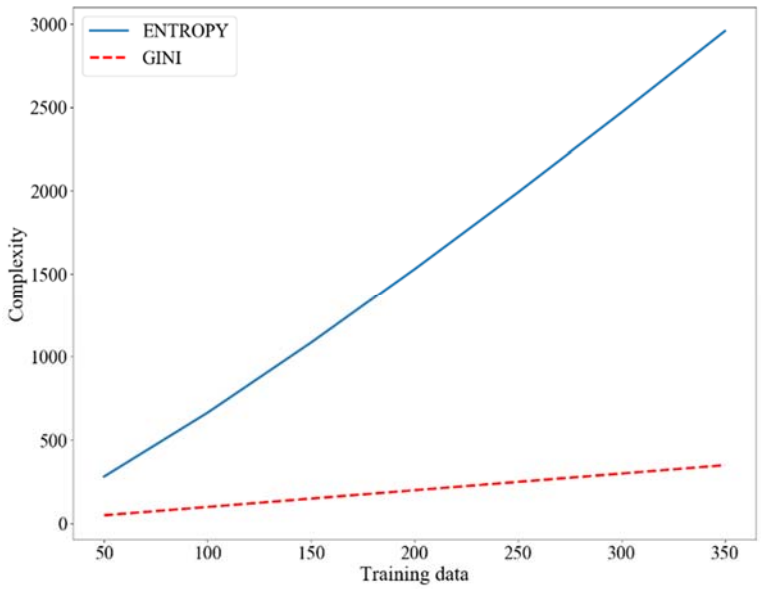

Figure 5. The computational complexities of information entropy and GINI impurity.

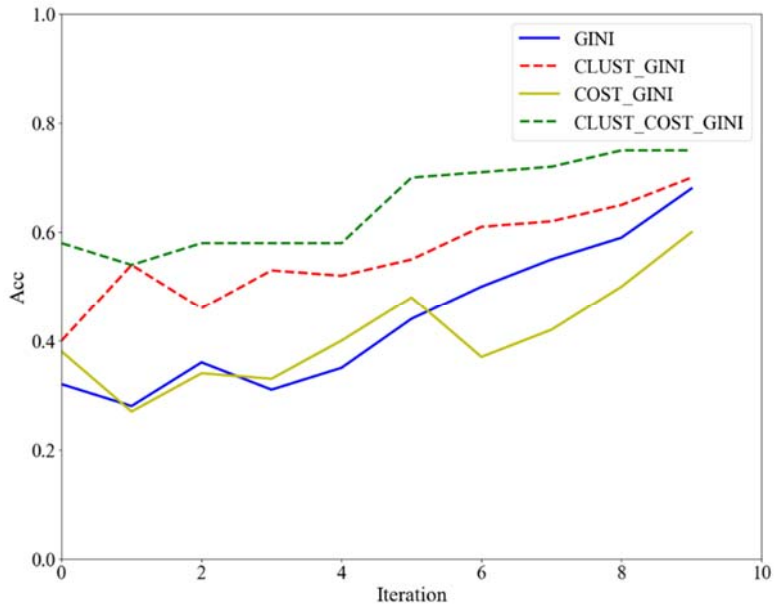

Figure 6. Model accuracies of different selection strategies.

4.4.3. Ablation experiment

This section aims to evaluate the combination strategy's performance based on the multi-standard proposed in this article. We conduct the ablation studies from three different aspects: (i) the effects on experimental results by adding cost or cluster selection strategies; (ii) the impacts of labeling time by adding cost selection strategy; (iii) the influence of experimental accuracy with β in combination strategy.

1) The impact of different selection strategies Table 6 and Figure 6 show the accuracies of varying selection strategies. Among them, GINI represents Gini impurity, CLUST indicates the combination of the K-means algorithm clustering, and COST stands for the labeling cost. Thus, CLUST_COST_GINI means the integration of K-means clustering, labeling cost, and Gini impurity.

According to the accuracy comparison in Table 5 and Figure 6, it can be found that the accuracy has been improved to a certain extent by adding the module CLUST to GINI, especially at the beginning of training. Interestingly, the accuracy decreases after increasing the COST based on GINI. It is because the model trades the accuracy for a low cost. By adding the modules CLUST and COST to GINI, the experimental performance achieves the best. It not only reduces the cost of labeling but also improves accuracy.

Table 6. Accuracies of different selection strategies.

Epochs Strategies	1	2	3	4	5	6	7	8	9	10
GINI	0.32	0.28	0.36	0.31	0.35	0.44	0.5	0.55	0.59	0.68
CLUST+GINI	0.4	0.54	0.46	0.53	0.52	0.55	0.61	0.62	0.65	0.7
COST+GINI	0.38	0.27	0.34	0.33	0.4	0.48	0.37	0.42	0.5	0.6
CLUST+ COST+GINI	0.58	0.54	0.58	0.58	0.58	0.7	0.71	0.72	0.75	0.75

2) The impact on labeling time. In the former experiment, we found that the accuracy of the uncertainty selection strategy considering the labeling cost in the later training process is lower than that on Gini impurity. It shows that adding the labeling cost factor is a negative impact on the model's training results. And it also proves that the selection strategy based on cost becomes the model accuracy's expense. Therefore, to verify whether the labeling cost can save labeling time, we asked the partner hospital experts to label the samples selected in the two cases (GINI and GINI+COST) and count the spending time. The comparison result is shown in Figure 7. Blue bars represent GINI, and red bars stand for GINI+COST.

It can be seen from the comparison that the labeling time with labeling cost (red bars) is lower than that without the labeling cost (blue bars), even though the gaps in labeling time of selected sentences are gradually becoming narrow with the gradual iteration of training. It is particularly apparent at the beginning of training.

3) The effect of different β values. To explore the influence of different β values from Eq (10) on NER's accuracy, this experiment conducted the test on the scale factor β in the combination strategy based on the uncertainty and the cost, aiming to select the most suitable.

By observing the curve in Figure 8, it is clear that β starts from 0.2 to 0.8, and the proportion of labeling costs also increases. When β rises to 0.8, the accuracy curve is quite different from the other three. If β is too small, the cost of labeling time is enormous, although the corresponding accuracy is relatively high. On the contrary, if β is too large, the cost of accuracy in exchange for low cost is too

great and meaningless. Therefore, the most suitable β value is 0.6, with the maximum proportion.

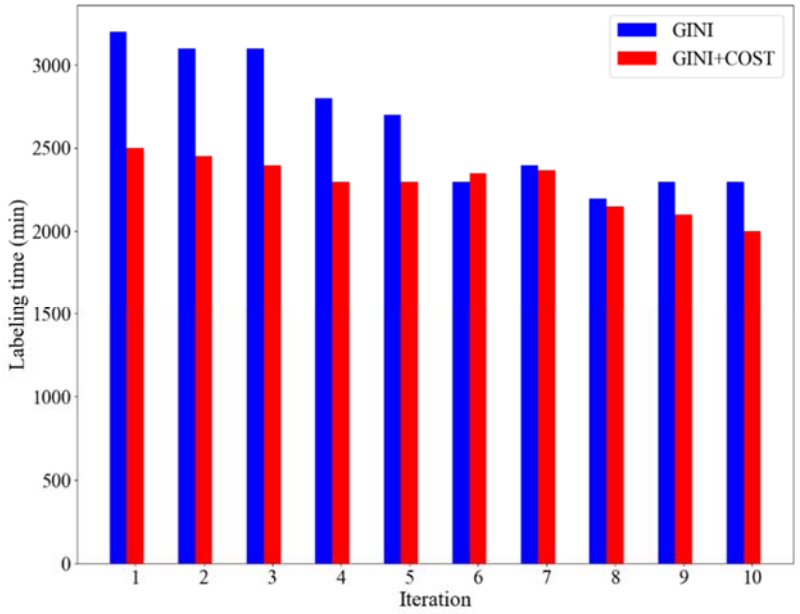

Figure 7. Time spent by experts in labeling samples.

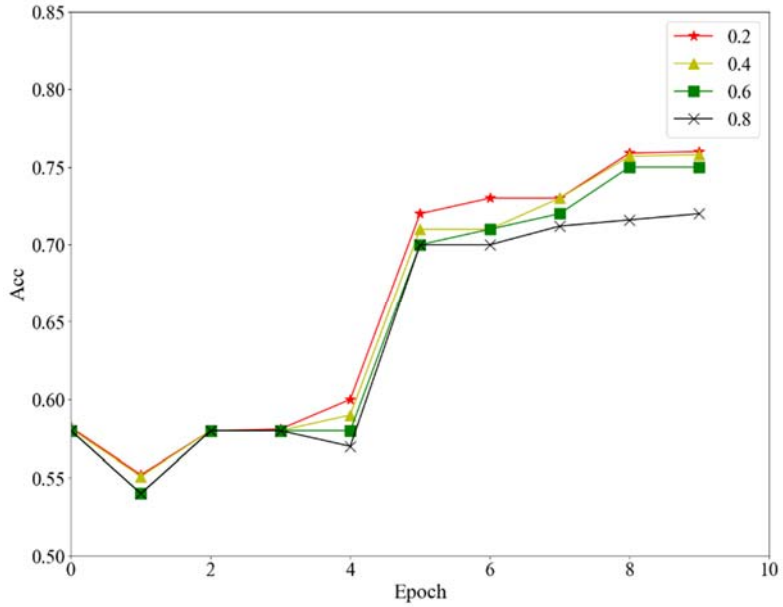

Figure 8. Accuracy curves under different β values.

4.4.4. Comprehensive experimental comparison

We aim to compare the performances of different selection strategy methods. To achieve the same performance accuracy, the number of training data from each selection strategy is different. Since the active learning method uses batch mode in this article, the number of training texts is proportional to training iterations–the lower the iteration, the better the performance. Therefore, we

compare the iteration numbers of different methods at the same accuracy level. All methods are tested within 100 iterations.

As shown in Figure 9, at the accuracy of 0.7, it only requires to iterate five times for the combination strategy (CLUST+GINI+COST), when GINI, RANDOM, and GINI+COST need 20, 22, and 25-times for iterations, respectively. Achieving the accuracy of 0.75 requires at least 12-times iterations for the combination strategy. In the meantime, GINI needs 24 times, RANDOM requires 36 times, and GINI+COST acquires 42 times. Besides, it can be observed that the combination strategy can reach an accuracy of 0.85 with 55 iterations. In comparison, the uncertainty selection strategy based on Gini impurity reaches the same accuracy level with 77 iterations in training. It should be noted that the RANDOM and COST are not included in the accuracy of 0.8 because the instability of RANDOM and the highest accuracy of COST is 0.78. We can conclude that the strategy based on uncertainty, labeling cost, and clustering has the best efficiency in improving model performance. Compared with the traditional supervised learning method of randomly selecting labeled data, the amount of labeled data is reduced by about 66.67%.

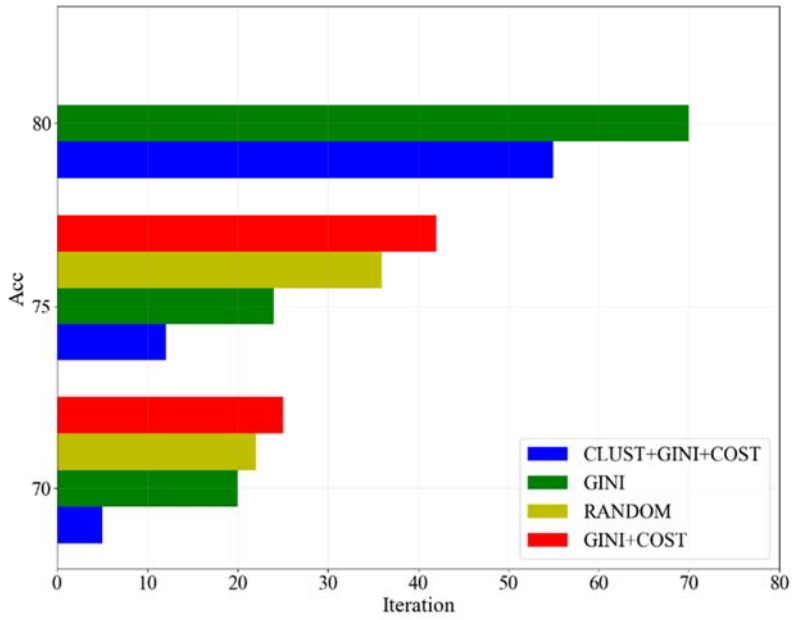

Figure 9. The number of iterations required for different performance levels.

4.4.5. Method improvement

We studied other possible solutions based on labeling cost in Eq (9). When the predicted result of a selected sentence is corrected, each word's modification time is assumed to be the same. In practice, the modification time of each word is different. Therefore, we provide another solution using the TF-IDF algorithm on the training sample to consider the word frequency of the corpus's text features. If the word occurs infrequently, it will get a longer labeling time. It is because these texts may be rare medical records or new medical records, so the experts need more time to label them.

The labeling time of the improved scheme and the original method is shown in Figure 10. It can be seen that the improved method always performs better than the original method. The possible

reason is that the original method's labeling cost is positively related to the sentence's length. It does not consider the memory ability of experts. With the improved method's help, similar samples only need to spend a small amount of time to complete the labeling of a sample.

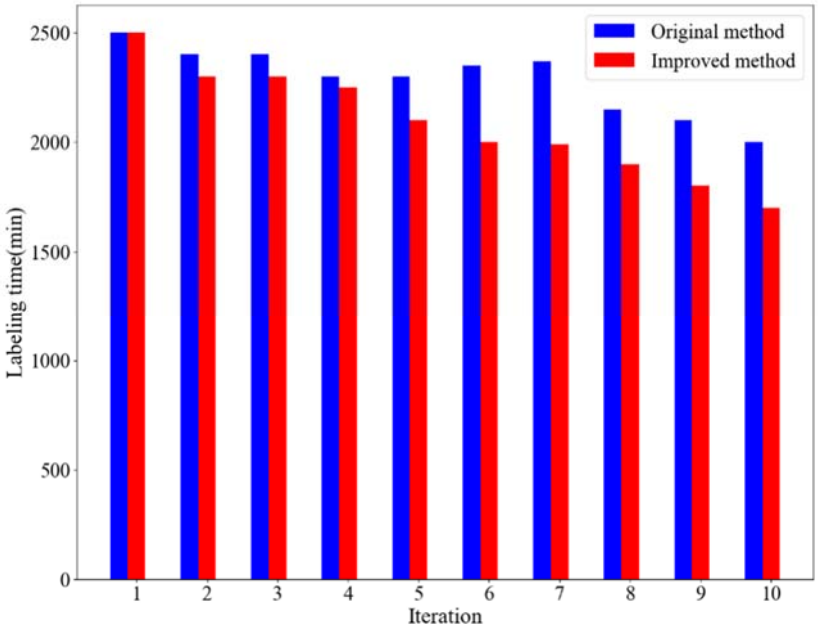

Figure 10. The improved cost-effective approach.

5. Conclusions

This paper proposed and implemented an entity recognition model of breast electronic medical records based on multi-criteria active learning. With the comparisons of different experiments, the proposed method's effectiveness and practicability are verified by effectively reducing the required labeling workload in practical applications. To solve the problems of brief sentences, semantic relevance, and sparse data in the medical text, the improved TF-IDF method based on Word2Vec is used to vectorize the text. Then the K-means algorithm is used to cluster the samples. After obtaining the clustering results, the uncertainty selection strategy based on Gini impurity is proposed to select the most useful labeling instances due to the high algorithm complexity and low accuracy of the traditional uncertainty selection strategy based on information entropy. The labeling cost is considered to reduce the labeling time for active learning in practical application scenarios. Finally, a novel active selection criterion is proposed, which balances between the uncertainty and the labeled cost. Compared with the traditional method of randomly selecting labeled data, the proposed method reduces the amount of data by 66.67%. Comprehensive experimental results show that the combination strategy can achieve optimal performance on a small amount of data after clustering.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2019YFE0190500.

Conflict of interests

No potential conflict of interest was reported by the authors.

References

1. C. Zeng, G. Hui, Construction of electronic medical record system for standardized diagnosis and treatment of breast cancer, *J. Chin. Med. Dev.*, **29** (2014), 46–48.
2. Q. M. Ling, Research on the advantages and development of electronic medical record in medical record management, *Electron. J. Gen. Stomatol.*, **7** (2020), 26–31.
3. L. Liu, D. B. Wang, Summary of research on named entity recognition, *J. Chin. Soc. Sci. Tech. Inf.*, **3** (2018), 329–340.
4. C. Y. Kun, L. T. A, M. Q. Zhu, D. J. C, X. Hua, A study of active learning methods for named entity recognition in clinical text, *J. Biomed. Inf.*, **58** (2015), 11–18.
5. Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, A. Anandkumar, Deep active learning for named entity recognition, preprint, arXiv:1707.05928.
6. W. W. Ning, L. Yang, G. M. Zu, L. X. Yan, Research progress of active learning algorithm based on sampling strategy, *J. Comput. Res. Dev.*, **49** (2012), 1162–1173.
7. W. R. Qi, L. X. Li, H. Y. Li, B. He, G. Yi, Research on active learning method for named entity recognition of Chinese electronic medical record, *Chin. Digital Med.*, **12** (2017), 51–53.
8. L. M. Qun, S. Martin, E. E. Khaled, M. B. A, Efficient active learning for electronic medical record de-identification, *AMIA Summits Transl. Sci. Proc.*, (2019), 462–471.
9. M. Kholghi, L. Sitbon, G. Zuccon, A. Nguyen, External knowledge and query strategies in active learning: a study in clinical information extraction, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, (2015), 143–152.
10. J. Zhu, E. H. Hovy, Active learning for word sense disambiguation with methods for addressing the class imbalance problem, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (2007), 783–790.
11. M. Bloodgood, K. Vijay–Shanker, Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets, preprint, arXiv:1409.4835.
12. K. Tomanek, U. Hahn, Reducing class imbalance during active learning for named entity annotation, in *Proceedings of the fifth international conference on Knowledge capture*, (2009), 105–112.
13. S. Ertekin, J. Huang, C. L. Giles, Active learning for class imbalance problem, in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, **1** (2007), 823–824.
14. S. Ertekin, J. Huang, L. Bottou, C. L. Giles, Learning on the border: Active learning in imbalanced data classification, in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, (2007), 127–136.
15. H. Guang, Z. C. Xia, H. X. Lei, A new SVM active learning algorithm and its application in obstacle detection, *J. Comput. Res. Dev.*, **46** (2009), 1934–1941.

16. B. C. Mei, Classification of weighted support vector machines based on active learning, *Comput. Eng. Des.*, **30** (2009), 966–970.
17. Y. F. Liang, *Research on Active Learning Algorithm Based on Expert Committee*, Master thesis, Ocean University of China, 2010.
18. L. Feng, *Research and Application of Active Semi-supervised K-means Clustering Algorithm*, Master thesis, Hebei University of Geosciences, 2018.
19. X. Li, Y. Guo, Adaptive active learning for image classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2013** (2013), 859–866.
20. M. Kholghi, L. D. Vine, L. Sitbon, G. Zuccon, A. Nguyen, Clinical information extraction using small data: an active learning approach based on sequence representations and word embeddings, *J. Assoc. Inf. Sci. Technol.*,**68** (2017), 2543–2556.
21. D. Angluin, Queries and concept learning, *Mach. Learn.*, **2** (1988), 319–342.
22. R. Grishman , B. Sundheim, Message understanding conference-6: A brief history, in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.
23. C. Friedman, P. O. Alderson, J. H. Austin, S. B. Johnson, A general natural-language text processor for clinical radiology, *J. Am. Med. Inf. Assoc.*, **1** (1994), 161–174.
24. W. S. Li, *Research on Chinese Electronic Medical Record of Named Entity Recognition Based on Improved Deep Belief Network*, Master thesis, Beijing University of Chemical Technology, 2018.
25. G. K. Savova, J. J. Masanz, P. V. Ogren, J. P. Zheng, S. W. Sohn, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, *J. Am. Med. Inf. Assoc.*, **17** (2010), 507–513.
26. S. T. Wu, H. F. Liu, D. C. Li, C. Tao, M. A. Musen, N. H. Shah, Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis, *J. Am. Med. Inf. Assoc.*, **19** (2012), 149–156.
27. E. F. Sang, F. D. Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, preprint, arXiv:0306050.
28. Y. Li, S. L. Gorman, N. Elhadad, Section classification in clinical notes using supervised hidden markov model, in *Proceedings of the 1st ACM International Health Informatics Symposium*, (2010), 744–750.
29. P. Y. Wang, D. H. Gi, Disease name extraction based on multi-label CRF, *Appl. Res. Comput.*, **1** (2017), 118–122.
30. F. Ye, Y. Y. Chen, G. G. Zhou, H. M. Li, Y. Li, Intelligent recognition of named entities in electronic medical records, *Chin. J. Biomed. Eng.*, **2** (2011), 98–104.
31. J. Liang, X. M. Xian, X. J. He, M. F. Xu, S. Dai, J. Y. Xin, A novel approach towards medical entity recognition in Chinese clinical text, *J. Healthcare Eng.*, (2017), 1–16.
32. G. Luo, X. Huang, C. Y, Z. Nie, Joint entity recognition and disambiguation, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (2015), 879–888.
33. A. Passos, V. Kumar, M. C. Andrew, Lexicon infused phrase embeddings for named entity resolution, preprint, arXiv:1404.5367.
34. Y. J. Zhang, Z. T. Xu, X. Y. Xue, A maximum entropy Chinese named entity recognition model of integrating multiple features, *J. Comput. Res. Dev.*, **6** (2008), 1004–1010.

35. A. Mccallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, *Comput. Sci. Dep. Fac. Publ. Ser. 11*, (2003), 188–191.
36. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, *J. Mach. Learn. Res.*, **12**(2011), 2493–2537.
37. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, preprint, arXiv:1603.01360.
38. A. Jagannatha, Y. Hong, Structured prediction models for RNN based sequence labeling in clinical text, in *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, (2016), 856–865.
39. J. Zhu, H. Wang, B. K. Tsou, M. Ma, Active learning with sampling by uncertainty and density for data annotations, *IEEE. Trans. Audio, Speech, Lang. Process.*, **18** (2012), 1323–1331.
40. X. Yan, *Research on Image Annotation Method Based on Active Learning*, Master thesis, Liaoning University of Technology, 2014.
41. L. Jin, Y. F. Cao, C. X. Su, J. Y. Ren, Multi-class image classification based on HS sample selection and BvSB feedback, *J. Guizhou Norm. Univ. (Nat. Sci.)*, (2014), 56–61.
42. Q. H. Zhao, *Two active learning methods*, Master thesis, He Bei University, 2010.
43. S. Ertekin, J. Huang, C. L. Giles, Active learning for class imbalance problem, in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (2007), 823–824.
44. H. S. Seung, M. Opper, H. Sompolinsky, Query by Committee, in *Proceedings of the fifth annual workshop on Computational learning theory*, (1992), 287–294.
45. D. D. Lewis, J. Catlett, Heterogeneous uncertainty sampling for supervised learning, in *Machine learning proceedings*, Morgan Kaufmann, 1994, 148–156.
46. T. Scheffer, C. Decomain, S. Wrobel, Active hidden markov models for information extraction, in *International Symposium on Intelligent Data Analysis*, Springer, Berlin, Heidelberg, (2001), 309–318.
47. S. Tong, D. Koller, Support vector machine active learning with applications to text classification, *J. Mach. Learn. Res.*, **2** (2002), 45–66.
48. A. Kapoor, E. Horvitz, S. Basu, Selective supervision: guiding supervised learning with decision-theoretic active learning, in *IJCAI*, **7** (2007), 877–882.
49. S. Arora, E. Nyberg, C. P. Rose, Estimating annotation cost for active learning in a multi-annotator environment, in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, (2009), 18–26.
50. J. Carroll, R. Haertel, P. McClanahan, E. K. Ringger, K. Seppi, Assessing the costs of sampling methods in active learning for annotation, *Fac. Publ.*, (2008), 185.
51. M. Kholghi, L. Sitbon, G. Zuccon, A. Nguyen, Active learning: a step towards automating medical concept extraction, *J. Am. Med. Inf. Assoc.*, **23** (2016), 289–296.
52. R. Q. Wang, X. L. Li, Y. L. Huang, B. He, Y. Guan, Research on active learning method of Chinese electronic medical record named entity recognition, *China Digital Med.*, **12** (2017), 51–53.
53. J. Z. Cheng, W. Qiang, A. Franklin, T. Cohen, H. Xu, Cost-sensitive active learning for phenotyping of electronic health records, *AMIA Summits Transl. Sci. Proc.*, **2019** (2019), 829–838.

54. Ö. Uzuner, B. R. South, S. Shen, S. L. Duvall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, *J. Am. Med. Inf. Assoc.*, **18** (2011), 552–556.
55. S. Pradhan, N. Elhadad, B. South, D. Martinez, L. Christensen, A. Vogel, Task 1: ShARe/CLEF ehealth evaluation lab 2013, in *CLEF (Working Notes)*, 2013.
56. G. S. Wang, X. J. Huang, Text classification model of convolutional neural network based on Word2vec and improved TF-IDF, *J. Chin. Mini-Micro Comput. Syst.*, **40** (2019), 210–216.
57. M. Kholghi, L. Sitbon, G. Zuccon, A. Nguyen, Active learning reduces annotation time for clinical concept extraction, *Int. J. Med. Inform.*,**106** (2017), 25–31.
58. T. H. Nguyen, A. Sil, G. Dinu, R. Florian, Toward mention detection robustness with recurrent neural networks, preprint, arXiv:1602.07749.
59. Z. H. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, preprint, arXiv:1508.01991.
60. Z. L. Yang, R. Salakhutdinov, W. Cohen, Multi-task cross-lingual sequence tagging from scratch, preprint, arXiv:1603.06270.

AIMS Press