Wang *et al. Journal of Cloud Computing: Advances, Systems and Applications* (2021) 10:4
https://doi.org/10.1186/s13677-020-00218-2

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT

Junshu Wang[1], Guoming Zhang[2,3], Wei Wang[4], Ka Zhang[1*] and Yehua Sheng[1,5]

Abstract

With the rapid development of hospital informatization and Internet medical service in recent years, most hospitals have launched online hospital appointment registration systems to remove patient queues and improve the efficiency of medical services. However, most of the patients lack professional medical knowledge and have no idea of how to choose department when registering. To instruct the patients to seek medical care and register effectively, we proposed CIDRS, an intelligent self-diagnosis and department recommendation framework based on Chinese medical Bidirectional Encoder Representations from Transformers (BERT) in the cloud computing environment. We also established a Chinese BERT model (CHMBERT) trained on a large-scale Chinese medical text corpus. This model was used to optimize self-diagnosis and department recommendation tasks. To solve the limited computing power of terminals, we deployed the proposed framework in a cloud computing environment based on container and micro-service technologies. Real-world medical datasets from hospitals were used in the experiments, and results showed that the proposed model was superior to the traditional deep learning models and other pre-trained language models in terms of performance.

Keywords: Cloud computing, Electronic medical record, BERT, Disease diagnosis

Introduction

Background

China is a country of large medical services and about 8 billion medical visits annually. Thus, hospitals in China, especially tertiary class hospitals, are always overcrowded with patients. This scenario directly leads to heavy workload for doctors, long queuing time, and poor medical treatment experience for patients. The appointment service has been applied in most hospitals to improve the efficiency of medical treatment and reduce the queuing time. Scheduling an appointment in advance through the website or mobile APP is convenient for patients. However, reservation registration also brings new problems. For instance, patients lacking professional medical knowledge have no idea how to choose the appropriate department of registration. Making an appointment registration according to previous experience is unsuitable because registering in an inappropriate department leads to patient inconvenience and tremendous waste in healthcare resources. Therefore, a method that predicts the type of disease according to patients' chief complaints and then accurately recommends registration departments for patients is the key to solve the problem.

With the widespread application of hospital information systems, abundant diagnosis and treatment data of patients have been collected by electronic medical record systems in hospitals. The rapid development of cloud computing, big data, and artificial intelligence has provided favorable conditions to construct self-diagnosis and registration department recommendation systems for patients with big medical data [1–3]. The medical data in electronic medical record systems are mainly

*Correspondence: zhangka81@126.com
[1]Key Laboratory for Virtual Geographic Environment Ministry of Education Nanjing Normal University, 210008 Nanjing, China
Full list of author information is available at the end of the article

unstructured text, which needs natural language processing (NLP) technology to model.

The task in this paper is mainly related to text classification. Text classification methods consist of traditional shallow algorithms (e.g., support vector machine (SVM), random forest, and Bayes) and deep learning algorithms (e.g., Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Hierarchical Attention Network (HAN), Bidirectional Encoder Representations from Transformers (BERT)). The deep learning method performs better than the traditional methods. In deep learning algorithms, the pre-trained language model BERT has been widely studied and applied, and has achieved state-of-the-art performance in many NLP tasks [4–6]. However, there are some limitations in applying the existing pre-training language model directly to medical text mining. First, the performance of BERT on medical text mining needs further evaluation because this model is trained on general text datasets. Second, medical and general texts differ in word distribution [7]. Training a BERT model in the medical domain to perform medical text mining is urgently needed. The number of parameters of BERT influences its performance. In general, more model parameters correspond to better effect. Large pre-trained models usually take up large space and run slowly in intelligent terminals with low computing power, causing difficulty for the models to run directly on terminals.

For the above mentioned challenges, we proposed an intelligent self-diagnosis and department recommendation model based on Chinese medical BERT in the cloud computing environment. The model was deployed in the cloud. It first predicted the type of disease on the basis of chief complaints, then recommended registration departments for patients, finally provided medical help-seeking advice for patients and avoided medical resources waste. In order to verify our model, a Chinese BERT model CHMBERT was trained on medical text data. The proposed framework CIDRS was tested on the inpatient dataset from a tertiary class hospital in Jiangsu Province, China. The experimental results showed that the proposed model achieved the best performance compared with the state-of-the-art methods.

Contributions

The contributions of this work are as follows:

1) A Chinese medical pre-trained language model trained on a large-scale medical text corpus from more than 100 hospitals was proposed. The text corpus was collected by the Jiangsu National Health Information Platform. The proposed model is the first medical BERT model trained on such a large scale of Chinese medical corpus, which is of great significance for evaluating the effectiveness of the pre-trained model in the Chinese medical domain.

2) An intelligent self-diagnosis and department recommendation model based on Chinese medical BERT (CHMBERT) was proposed. This model predicted diseases and recommended registration departments according to chief complaint. Thus, it can provide medical help-seeking advice for patients effectively and avoid medical resources waste. Furthermore, the model was deployed in the cloud computing framework, which can solve the problem of low terminal computing power.

3) Experiments were performed on a large-scale medical dataset by comparing with current popular algorithms and other pre-trained models to evaluate the effectiveness of our proposed framework.

The rest of this paper is organized as follows. In "Preliminary knowledge" section, we depict the pre-trained model BERT and medical texts. "Methodology" section introduces the proposed self-diagnosis and department recommendation model and the corresponding cloud computing framework CIDRS. "Experiments" section evaluates the proposed CHMBERT model over real-world medical texts. "Related work" section summarizes the current research. Finally, we conclude our study and provides future research direction in "Conclusion" section.

Preliminary knowledge

This section discusses the preliminary knowledge of the medical text data and pre-trained BERT model.

Medical text data

Medical text data mainly include patient demographic information, past history data, clinical diagnosis and treatment data, and so on [8].

Demographic information: contains the patient's basic information, including name, gender, age, and address.

Past history data: contains summary information of the patient's historical health, including disease history, allergy history, surgical history, trauma history, blood transfusion history, family/genetic history, and hospital history.

Clinical diagnosis and treatment data: contains the patient's detailed clinical diagnosis and treatment information, including the symptoms and signs, chief complaint, history of present illness, diagnosis name, treatment process description, test results, and examination report.

BERT

BERT [4] is a pre-trained language model that has been widely used in the past year. It has achieved state-of-the-art effects in multiple downstream NLP tasks, such as Machine Translation, Named Entity Recognition, Text

Classification, Reading Comprehension, and Question Answering.

BERT extracts context features by using a bidirectional transformer encoder [9], which has deeper levels and better parallelism. In the pre-training process of BERT, the input is constructed by summing over Token Embedding, Segment Embedding, and Position Embedding. New target tasks, namely, Masked Language Model and Next Sentence Prediction, are designed to analyze the relationship between words and sentences and to learn the corresponding expressions. The pre-trained BERT model can be fine-tuned through an additional output layer, which is widely applicable to the construction of NLP downstream tasks without large architectural modifications for specific tasks. Due to the space limitation, the detailed introduction of BERT can be found in the reference [4].

Methodology

Our proposed CIDRS framework includes a cloud computing platform and a self-diagnosis and department recommendation model. The model is deployed on the cloud computing platform after executing offline training on GPU. The users upload chief complaints to the cloud platform through terminals, and then the deployed model predict disease and recommend departments for users.

Cloud computing framework

The self-diagnosis and department recommendation model is hard to deploy on terminals because of its high computational power and large storage space requirements. Cloud computing service can solve the low computational power problem of user terminals [10, 11]. By adopting cloud computing architecture, the proposed model can be efficiently run on multi-CPUs/GPUs computing resources, whose capacity can be expanded dynamically as users' requests increase. It reduces the cost of hardware and maintenance for service providers [12, 13].

In this study, the cloud computing framework was deployed based on container and micro-service technologies. All of the containers were deployed on a Kubernetes cluster. Figure 1 illustrates the proposed cloud computing framework, which mainly includes Web application service, authentication service, query processing service, model service, monitoring service, a configuration database, and a user database. Web application service served the middle layer, which was called by the user terminals by publishing a REST service interface. Considering the privacy of medical diagnosis and treatment data, we executed bidirectional authentication on the client and the server by using a Kerberos authentication service to guarantee the reliability of both communication sides and the security of data transmission. Request data were

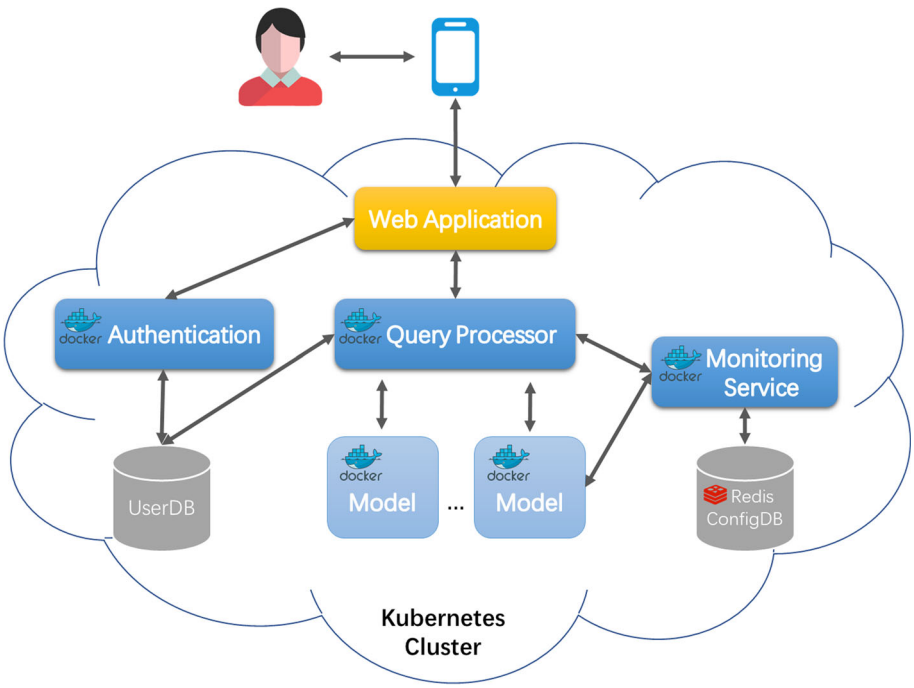


Fig. 1 The cloud computing architecture of proposed model

listened by query processing service and routed to the model service deployed on the Kubernetes cluster, which included a disease prediction model and a department recommendation model. The model service predicted the type of disease and recommended a department on the basis of the input request data. The monitoring service was responsible for monitoring latency and workload throughput of the query processing service and the model service. Adding copies of the model service is convenient when its single instance fails to meet the throughput requirements of the service workload. The user and configuration databases were used to store user information and configuration parameters for service components, respectively. Our proposed cloud computing framework guaranteed the flexibility and security of the whole system.

The above framework was implemented based on Java and TensorFlow Serving. The Java-based Web application service, authentication service, query processing service, model service, and monitoring service were developed with SpringBoot, and corresponding docker images were created based on CentOS base image and JDK1.8. The docker image of the user database was created with the MariaDB base image, and the docker image of the configuration database was created with the Redis base image. The docker image of the model service was built with the base image of TensorFlow Serving, and nvidia-docker should be installed on the hosts if GPU was used. When deploying the framework in production environment, at least three hosts are required to form a Kubernetes cluster. It is recommended to configure 64GB memory, 32 core CPU, 4T disk and 1 Tesla V100 graphics card for each host. All the services are deployed via containerization in a Kubernetes cluster. In order to guarantee high availability, each service needs to start at least two instances. Containers are spread across hosts in Kubernetes, therefore, when one host fails, the services on other hosts can still run normally.

Disease prediction and department recommendation model

A chief complaint [14] includes a patient's self-reported symptoms, signs, nature, and duration. The model in this section makes predictions according to the chief complaint of a patient, and outputs the possible disease category and the recommended department. The disease prediction and department recommendation task can be transformed into two separate text classification problems: 1) predicting the disease category according to the chief complaint and 2) predicting the department category according to the chief complaint. Considering the great success of the BERT-based pre-trained model in NLP tasks, we pre-trained a Chinese medical BERT model and obtained two fine-tuned models by fine-tuning the classification tasks on disease prediction and department recommendation.

Pre-training a Chinese medical BERT model

The traditional Chinese BERT model is a universal language representation model pre-trained on Wikipedia corpus. However, medical texts contain several professional terms and differ in word distribution from general texts. Therefore, NLP models designed for universal natural language understanding always perform poorly in medical text mining tasks [7]. To solve this problem, we constructed a medical text corpus based on more than 100 hospitals from the Jiangsu Regional Health Information Platform, including past history data and clinical diagnosis and treatment data. The corpus data, which consist of complaints, hospital admissions, progress notes, and discharge records, were obtained from the electronic medical record system in hospitals. The data were about 185GB in size. CHMBERT, a BERT model that focuses on the Chinese medical domain, was trained on this corpus. The original BERT code, model structure, and parameters were used to train the CHMBERT model on the Chinese medical corpus. The original Chinese BERT model parameters were utilized for CHMBERT initialization instead of training from scratch to improve the computational efficiency. In the pre-training process, the maximum length of the sentence was set to 128, and the number of training steps was set to 10 million (100K) steps. Finally, about a month was needed to complete the pre-training process on a Tesla V100 GPU.

Fine-tuning CHMBERT for disease prediction and department recommendation

On the basis of the pre-trained CHMBERT model, a fully connected output layer *Softmax* was used to fine-tune the two classification tasks in this paper. The input text sequence of a chief complaint was described as $X = \{x_1, x_2, \ldots, x_L\}$, and x_i denotes a Chinese character, $0 < i \leq L$, where L is the maximum length of input text sequence. For example, the input chief complaint is '发热咳嗽一天', then X ={'发','热','咳','嗽','一','天'}. The text sequence X was encoded into a fixed-length sentence vector S through the CHMBERT model, which was expressed by Formula (1) as follows:

$$S = CHMBERTsent(X) \tag{1}$$

where $CHMBERTsent(.)$ represents the transformation from text sequence into sentence vector. Then the sentence vector S was passed into a fully connected layer with *dropout* using Formula (2) as follows:

$$f_c = \text{Relu}\,(W \cdot S + b_s) \tag{2}$$

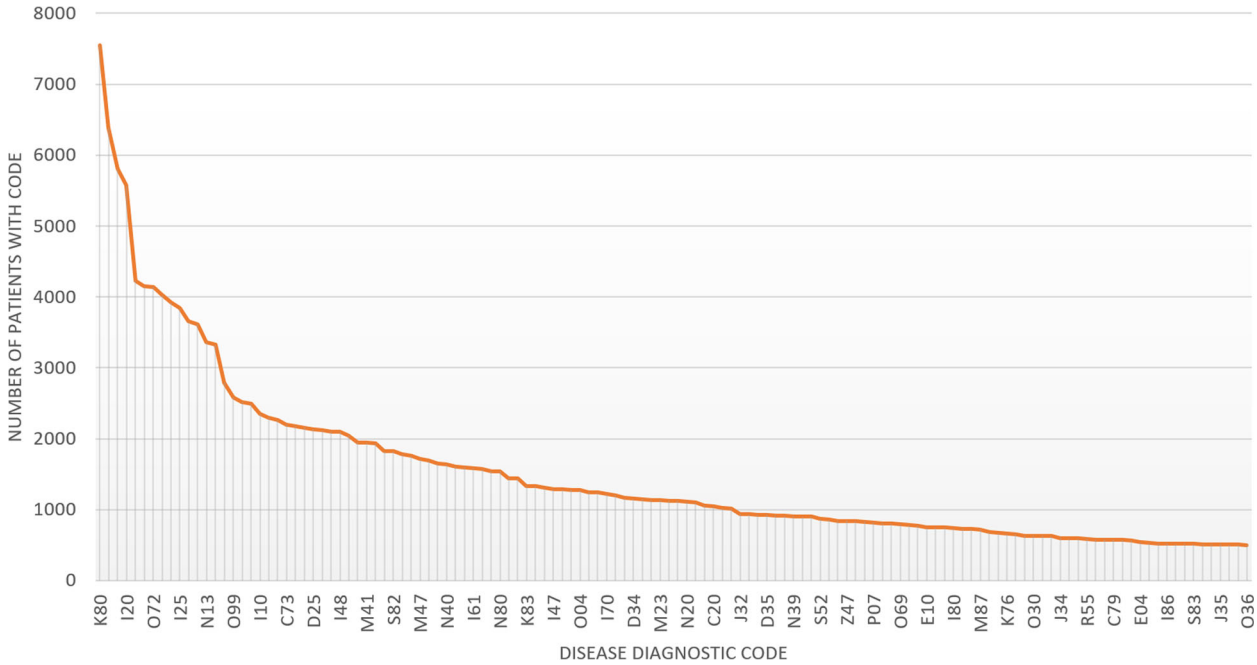


Fig. 2 The distribution of the number of patients with 130 kinds of diseases

Finally, the *Softmax* layer will output the probability distribution of disease or department category according to Formula (3) as follows:

$$p_k = \frac{\exp\left(w_k^T S\right)}{\sum_j \exp\left(w_k^T S\right)} \tag{3}$$

where p_k denotes the probability that sentence vector S belongs to category k, and $\sum_j \exp\left(w_k^T S\right)$ is a normalized item. The *cross-entropy* loss function and *Adam* optimization algorithm were used for fine-tuning the model parameters.

Experiments

Dataset

The purpose of our model was to predict disease category and recommend registration department according to patient's chief complaint. To verify the performance of our model, we selected 200,000 inpatients' chief complaints and the corresponding disease diagnosis codes and treatment departments from a tertiary class hospital from January 2015 to December 2018. In electronic medical records, the disease diagnosis codes were classified by International Classification of Diseases (ICD)-10 (https://www.who.int/classifications/icd/en/), and only the first three bits of the ICD code were considered in this paper. After data cleaning and filtering, 198,000 records were collected, including 130 types of disease diagnosis and 25 departments, covering about 80% of the inpatients. In the dataset, the maximum and minimum lengths of sentences were 36 and 2, respectively, and the average length was 12. The total number of Chinese characters was 1456. The dataset was divided into a training set, a validation set, and a test set in a ratio of 70:15:15.

Figure 2 illustrates the distribution of the number of patients with 130 types of diseases, which follows the power-law distribution. The top 30 diseases account for about 50% of patients. The disease with the largest number of patients is K80 (Cholelithiasis), and the disease with the least number of patients is O36 (Maternal care for other known or suspected fetal problems). Figure 3 illustrates the distribution of the number of patients in 25 departments, which presents the power-law distribution as well, and the top three departments (general surgery department, obstetrics department and Vasculocardiology Department) in the number of patients accounted for about 35% of the total patients.

Baselines

- TextCNN [15]. It is a text classification algorithm based on CNN. It utilizes multiple convolution kernels in different sizes to extract key information from sentences, which can capture local correlation of sentences. TextCNN has simple architecture and fast training speed, achieving state-of-the-art results on multiple datasets.
- BiLSTM [16]. RNN is a widely applied NLP model that can process variable length text sequences and learn long distance dependencies from sentences. In this experiment, a single-layer bidirectional LSTM network was utilized to classify the input text.

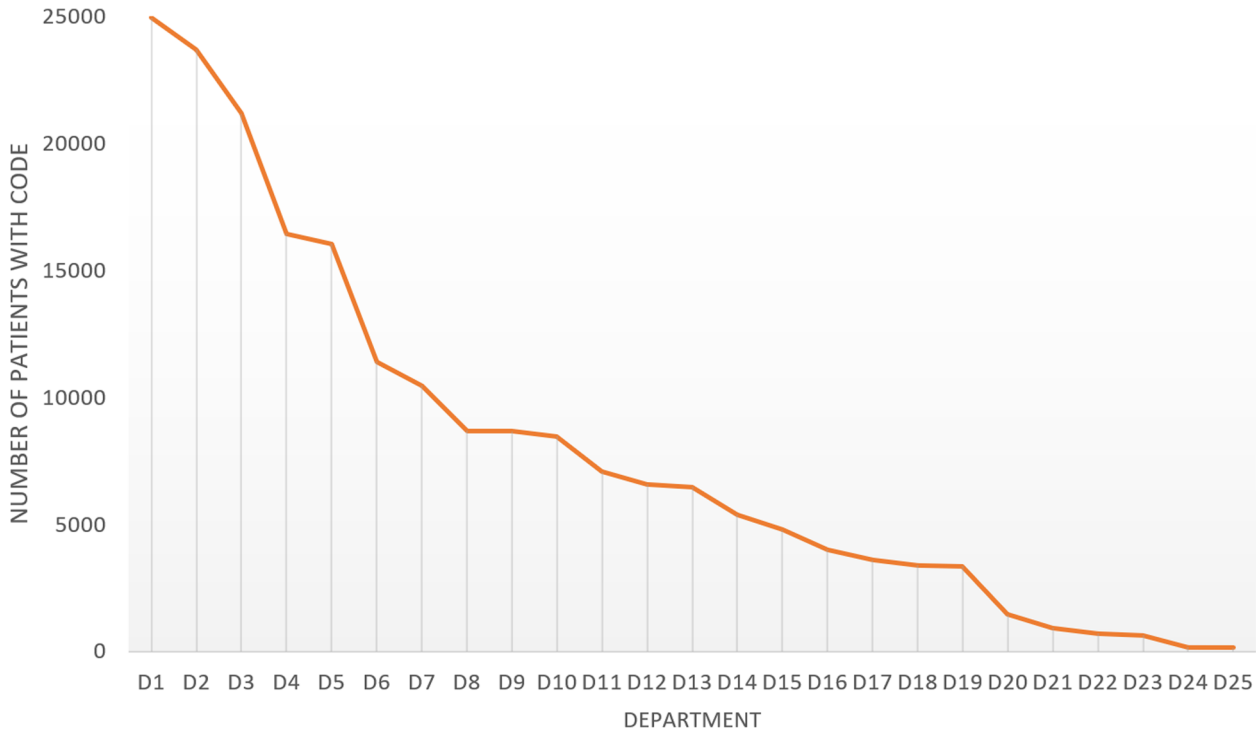


Fig. 3 The distribution of the number of patients in 25 departments

- LEAM [17].It is a model based on attention mechanism. It performs well in text representation by learning the joint embedding of word and label in the same space. Compared with other attention-based models, LEAM needs fewer model parameters and converges faster, and has good interpretability.
- Transformer [9].It is a sequence processing model based on self-attention mechanism, which can learn long-distance dependency from sentences. It can run in parallel paradigm and is the basis of BERT and other pre-trained models.
- BERT-base [4]. It is the original Chinese BERT pre-trained model published by Google, which achieves the state-of-the-art performance in many text classification tasks.
- BERT-wwm [18]. The updated version of BERT, published by Harbin Industrial University, is a Chinese pre-trained model based on Whole Word Masking technology. Its performance is slightly better than that of the original BERT in sentence classification task.

Implementation details

We selected the optimum parameters on the validation dataset through parameter tuning. The differences of the experimental results with different parameters were small, indicating that the clinical dataset in this paper was insensitive to parameters. In addition, Chinese BERT-base was segmented by character size without considering Chinese word segmentation in traditional NLP. In our experiments, the word segmentation was not under consideration either.

The same word embedding size, batch size, and maximum sentence length of 64, 128, and 36, respectively, were adopted in the models of TextCNN, LSTM, LEAM, and Transformer. The Adam algorithm was utilized for optimization. The number of iterations (epochs) was not limited, and the training process was conducted until the accuracy was not improved for 10 consecutive iterations. The parameters were set as follows:

TextCNN: Four types of convolution kernels with sizes of 2, 3, 4, and 5 were used. Each convolution kernel contained 128 kernels. The fully connected layer contained 256 neurons. The dropout was 0.5, and the learning rate was 1e4.

BiLSTM: The number of neurons in the LSTM hidden layer and the full connection layer was 128, dropout was 0.2, and the learning rate was 0.001.

LEAM: The label penalty coefficient was 1.0, the convolution kernel size was 3, and the number of neurons in the hidden layer was 300. The dropout was 0.5, and the learning rate was 0.001.

Transformer: The numbers of encoder layers and heads were 4 and 8, respectively, and the number of neurons in the Point wise feed forward network was 512. The dropout was 0.1, and the learning rate was 2e5.

BERT-base: The parameter setting should be same as that in the original BERT model when tuning the pre-trained model. The parameters in this paper were set as follows. The maximum sentence length was 36, and the batch size was 16. The number of epochs ranged from 1 to 5, and the tuning ranges of learning rates were 5e-6, 1e-5, 2e-5, 3e5, 4e5, and 5e-5 [4].

The parameter setting and the corresponding tuning ranges of BERT-wwm and CHMBERT were the same as those in BERT-base.

Experimental results

The commonly used Accuracy and F1 score in NLP classification task were used as evaluation criteria to compare the effects of different models. The same chief complaints may lead to different diseases; for instance, stomachache may be caused by enteritis, appendicitis, or other diseases. Therefore, top-k prediction results were calculated when predicting the type of disease. The k values were set to 1, 5, and 10, respectively, in these experiments. Similarly, more than one choice of first diagnosis department may be present on the basis of chief complaints. Thus, we obtained the prediction results of top-k when k=1, 2, and 3 when predicting the departments. The experimental results of disease and department prediction of different models are shown in Tables 1 and 2.

Tables 1 and 2 show that the pre-trained models based on BERT were significantly better than other state-of-the-art models. The CHMBERT model proposed in this paper performed the best among the tested models, which indicated that the pre-trained model had great potential in medical NLP task. As for the non-pre-trained models, text-CNN performed the best, followed by the Transformer models, whereas LSTM and LEAM performed the worst.

In the disease prediction experiment, the proposed CHMBERT model showed obvious advantages in the top-1 prediction. Compared with those of the sub-optimal model BERT-wwm, the accuracy and F1 of CHMBERT improved by 0.16% and 0.39%, respectively. Compared with those of text-CNN, which performed the best among the non-pre-trained models, the accuracy and F1 of CHMBERT improved by 0.9% and 1.35% respectively. In the prediction of top-5 and top-10, the performance of CHMBERT was similar to that of the sub-optimal model and slightly better than that of the text-CNN model.

In the department prediction experiment, our CHMBERT model achieved the best results. Compared with those of the sub-optimal model, the accuracy and F1 of CHMBERT improved by 0.14% and 0.59%, respectively, in the top-1 prediction. Compared with those of text-CNN, the prediction accuracy and F1 of CHMBERT improved by 0.79% and 1.74%, respectively, in the top-1 prediction. In the prediction of top-2 and top-3, the CHMBERT model also performed better than the BERT-wwm and text-CNN models.

Parameters discussion

We compared the performance of the CHMBERT model in disease prediction and department prediction with different learning rates and epochs. Figure 4 shows the top-1 prediction accuracy with different epochs when the learning rate was fixed with 2e5. Figure 5 shows the top-1 prediction accuracy with different learning rates when the number of epochs was set to 3.

As shown in Figs. 4 and 5, the prediction accuracy of CHMBERT was less affected by parameters. When the learning rate was fixed at 2e5 and the number of epochs varied from 1 to 5, the differences between the maximum and minimum values of disease prediction accuracy and department prediction accuracy were 1.58% and 0.78%, respectively. When the number of epochs was fixed at 3, the differences between the maximum and minimum values of the disease prediction accuracy and department prediction accuracy were 1.11% and 0.51% under different

Table 1 Average macro accuracy and F1-score for disease prediction

Methods	Average macro (Top-1)		Average macro (Top-5)		Average macro (Top-10)	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
TextCNN	65.38	60.6	92.91	91.49	97.27	96.73
LSTM	64.18	59.68	91.9	90.08	96.61	95.87
LEAM	63.44	55.79	92.1	90.2	96.9	96.3
Transformer	65.11	59.97	92.74	91.24	97.12	96.6
BERT-base	65.95	61.36	**93.11**	**91.59**	97.28	96.78
BERT-wwm	66.12	61.56	93.06	91.47	**97.34**	96.82
CHMBERT	**66.28**	**61.95**	93.08	91.58	97.27	**96.83**

The best performance is boldfaced

Table 2 Average macro accuracy and F1-score for disease prediction

Methods	Average macro (Top-1)		Average macro (Top-2)		Average macro (Top-3)	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
TextCNN	85.87	72.32	94.74	84.52	97.16	89.77
LSTM	84.72	70.98	94.05	83.46	96.55	89.00
LEAM	84.64	68.35	94.02	83.04	96.73	88.88
Transformer	84.98	69.59	94.37	83.87	96.95	89.51
BERT-base	86.47	73.36	95.04	85.29	97.32	89.62
BERT-wwm	86.52	73.47	95.03	84.47	97.24	89.73
CHMBERT	**86.66**	**74.06**	**95.18**	**85.30**	**97.44**	**90.67**

The best performance is boldfaced

learning rates, respectively. In general, when the number of epochs was small (such as 1, 2) and the learning rate was small (such as 5e6, 1e5), the performance was poor. These results indicate that 3 or 4 is the recommended number of epochs while 2e5, 3e5, or 5e5 is the recommended learning rate.

Related work

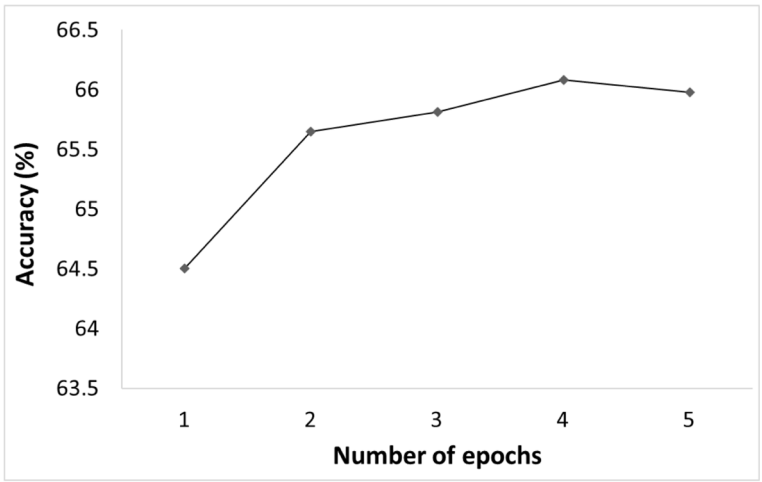
Text classification and pre-trained model

Text classification has attracted considerable attention as an important NLP task. In the early stage, shallow machine learning models (e.g., SVM [19] and logistic regression [20]) were utilized for text classification, and the performance was highly dependent on manually extracted features. With the rapid development in recent years, deep learning models that can automatically extract text features have been widely used in NLP tasks and achieved optimal results in text classification tasks, such as CNN, RNN, and the variants, such as GRU and LSTM/bi-lstm. The TextCNN [15] proposed by Kim et al. utilizes multiple convolution kernels in different sizes to capture the local features of sentences. The deep pyramid CNN [21] proposed by Johnson et al. extracts long-distance text dependencies through a deeper CNN network for text classification. Lai et al. proposed a Recurrent CNN [22] model to capture text context information through Bidirectional RNN and extract the features that are important to text classification via CNN. Considering that each word in a sentence has different levels of importance in classification, Yang et al. proposed a HAN [23] model. The text was represented by the hierarchical structure of "word-sentence-article," and different weights were assigned to words and sentences according to the attention mechanism, which can effectively improve the long text classification accuracy.

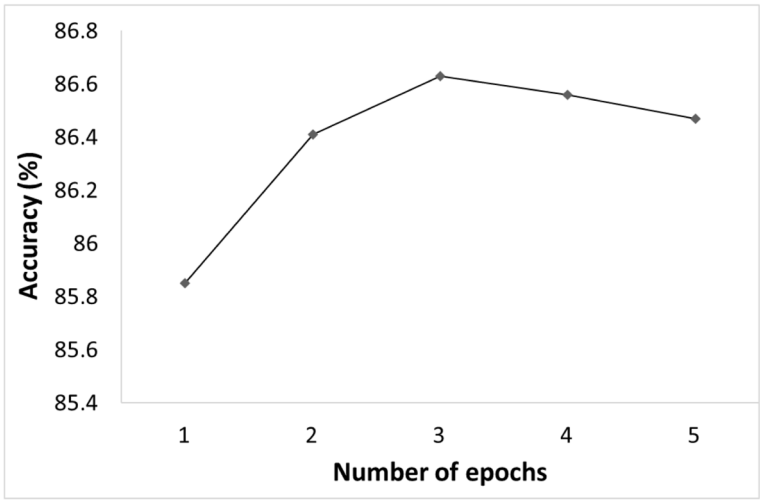
In the past year, the pre-trained models represented by BERT, such as XLNet [5] and RoBERTa [6] , have achieved great success in many NLP tasks. In terms of Chinese pre-trained models, ERNIE [24] as a representation model of knowledge enhancement was proposed by Baidu Company. It exceeded BERT when applied in Chinese datasets. The Chinese pre-trained model BERT-wwm [18] with full word coverage was released by Harbin Institute of Technology. It demonstrated the best performance among current Chinese pre-trained models. As for domain-oriented pre-trained models, Beltagy [25] et al. trained SciBERT for the scientific domain based on the scientific publication corpus, and it performed well in the text analysis of scientific datasets. Lee et al. proposed the BioBERT [7] model in the biomedical domain and found that it can achieve the best effect on multiple biomedical datasets. The lack of a pre-trained model in the Chinese medical domain limits the application of pre-trained models in medical text mining.

Disease classification

Several studies on automatic disease classification have been published. Most recent research has focused on deep neural network model. Shi [26] et al. used the LSTM network in character and word level to learn the diagnosis description and the implied representation of ICD name and matched them through the attention mechanism to achieve the automatic coding of ICD. Mullenbach [27] et al. proposed a model to predict disease diagnosis codes according to patients' discharge records based on CNN and label attention mechanism, which made the model highly interpretable. Zeng [28] proposed a deep transfer model that transfers the knowledge learned from the Medical Subject Headings index task to the ICD coding task and improved the effect of ICD coding. Li[29] et al. learned text patterns in different lengths based on a convolutional layer with multiple filters, and augmented the acceptable domain through the residual convolutional layer to classify ICD, which achieved good performance in MIMIC medical datasets. At present, disease diagnosis classification mainly focuses on English datasets, but no Chinese medical data related research has been found in this area. In addition, research on disease classification using BERT and other pre-trained models is lacking.



(a) Disease Prediction

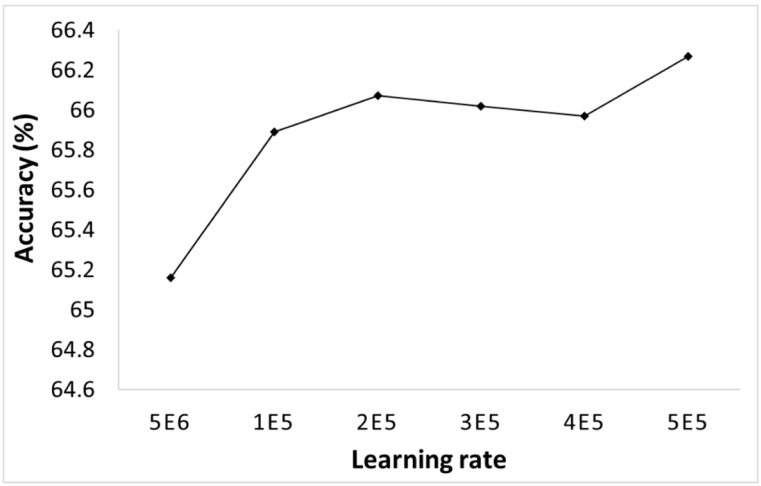


(b) Department Prediction

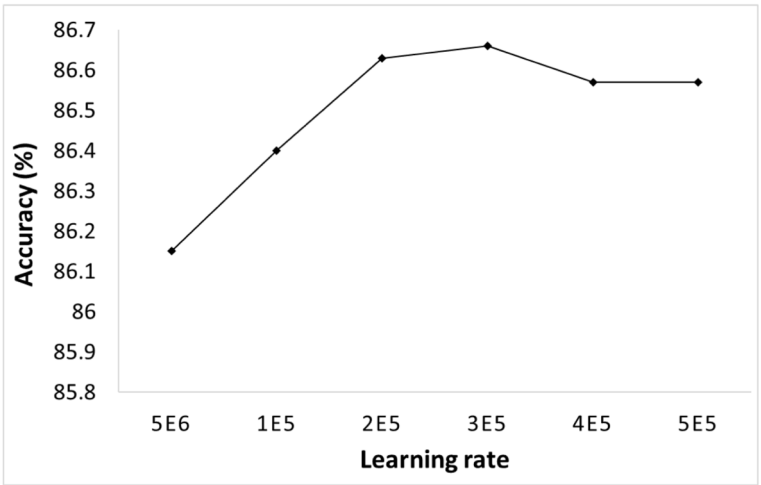
Fig. 4 The prediction accuracy with different epochs when the learning rate was 2e5

Cloud computing

Cloud computing has the advantages of virtualization, high availability and scalability, and low requirements for the users' terminals. It can be quickly deployed and has been widely applied in different domains [30–32]. Many issues remain to be tackled in cloud computing. Recent studies have focused on the security issues of cloud computing platforms and their application services [33, 34], privacy protection [35–37], optimization of energy consumption [38, 39], load balancing and resource scheduling [40–42]. In terms of cloud deployment for machine learning models, some frameworks have been proposed, such as Clipper (http://clipper.ai/), developed by UC Berkeley RISE Lab and Graphpipe (https://oracle.github.io/graphpipe/), which is the cloud deployment tool of Oracle's open source machine learning model. However, the cloud deployment of machine learning models remains to have many challenges, such as model extension and scalability, performance tuning, security, continuous integration, and deployment, which need further study. In terms of disease diagnosis based on cloud computing, Chen et al. [43] proposed a Disease Diagnosis and Treatment Recommendation System (DDTRS) based on Apache Spark cloud platform, which has high performance and low latency response. Lin et al. [8] put forward a cloud-based framework for implementing Home-diagnosis. In the framework, a distributed Lucene-based search engine was designed to provide scalable online and highly concurrent medical record retrieval service.



(a) Disease Prediction



(b) Department Prediction

Fig. 5 The prediction accuracy with different learning rates when the number of iterations was 3

Conclusion

A cloud computing service framework for disease prediction and department recommendation was proposed to guide patients to seek medical diagnosis and treatment effectively and avoid the waste of medical resources. A pre-trained language model in the Chinese medical domain CHMBERT was trained on large-scale Chinese medical corpus for the first time and used to optimize disease prediction and department recommendation tasks. Experimental results on the real-world medical datasets showed that our model achieved the best effect, which was superior to the traditional deep learning models and other pre-trained models. The pre-trained model for the medical domain has great potential in medical text mining tasks. In addition, our model provided services through the cloud computing environment, which can overcome the insufficient computing power of user terminals.

In our future work, we will utilize additional medical data to train our model for disease prediction and department recommendation and further improve the performance and availability of the model. In addition, we will further optimize the pre-trained model in the medical domain and try additional parameters and other advanced pre-trained methods, such as RoBERTa, XLNet, and ALBERT. We will further evaluate the performance of CHMBERT in other Chinese medical text mining tasks.

Abbreviations

EMR: Electronic medical record; BERT: Bidirectional encoder representations from transformers; CNN: Convolutional neural networks; LSTM: Long short-term memory; RNN: Recurrent neural network; GRU: Gated recurrent unit; NLP: Natural language processing; ICD: International classification of diseases

Authors' contributions
Junshu Wang, Guoming Zhang, Wei Wang, Ka Zhang and Yehua Sheng conceived and designed the study. Guoming Zhang and Wei Wang performed the simulations. Junshu Wang prepared the original draft. Ka Zhang and Yehua Sheng reviewed and edited the paper. All authors reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding
This work is supported in part by National Natural Science Foundation of China under Grant 41631175, the Natural Science Foundation of Jiangsu Province under Grant No. BK20171037, the Natural Science Foundation of Jiangsu Province under Grant No. BK20201372, the Program of Natural Science Research of Jiangsu colleges and universities under Grant No.17KJB170010.

Availability of data and materials
The data supporting the conclusions of these findings cannot be shared at this time as the data also forms part of an ongoing study.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Key Laboratory for Virtual Geographic Environment Ministry of Education Nanjing Normal University, 210008 Nanjing, China. [2]Department of Computer Science and Technology, Nanjing University, 210023 Nanjing, China. [3]Health Statistics and Information Center of Jiangsu Province, 210008 Nanjing, China. [4]Tencent Technology (Shenzhen) Co., Ltd, 518100 Shenzhen, China. [5]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, 210023 Nanjing, China.

Received: 8 September 2020 Accepted: 2 December 2020
Published online: 15 January 2021

References

1. Qi L, He Q, Chen F, Dou W, Wan S, Zhang X, Xu X (2019) Finding all you need: web apis recommendation in web of things through keywords search. IEEE Trans Comput Soc Syst 6(5):1063–1072
2. Zhang S, Choo K-KR, Liu Q, Wang G (2018) Enhancing privacy through uniform grid and caching in location-based services. Futur Gener Comput Syst 86:881–892
3. Liu H, Kou H, Yan C, Qi L (2020) Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph. Complexity 2020:1–15
4. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. ACL. pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
5. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems. MIT Press. pp 5754–5764. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf
6. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
8. Lin W, Dou W, Zhou Z, Liu C (2015) A cloud-based framework for home-diagnosis service over big medical data. J Syst Softw 102:192–206
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł., Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems. MIT Press. pp 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
10. Xu X, Mo R, Dai F, Lin W, Wan S, Dou W (2019) Dynamic resource provisioning with fault tolerance for data-intensive meteorological workflows in cloud. IEEE Trans Ind Inform. https://doi.org/10.1109/TII.2019.2959258
11. Wan S, Li X, Xue Y, Lin W, Xu X (2020) Efficient computation offloading for internet of vehicles in edge computing-assisted 5g networks. J Supercomput 76(4):2518–2547
12. Wang S, Zhou A, Bao R, Chou W, Yau SS (2018) Towards green service composition approach in the cloud. IEEE Trans Serv Comput. https://doi.org/10.1109/TSC.2018.2868356
13. Zhou J, Hu XS, Ma Y, Sun J, Wei T, Hu S (2019) Improving availability of multicore real-time systems suffering both permanent and transient faults. IEEE Trans Comput 68(12):1785–1801
14. Wagner MM, Hogan WR, Chapman WW, Gesteland PH (2006) Chief complaints and icd codes. Handb Biosurveillance:333–359. https://doi.org/10.1016/b978-012369378-5/50025-9
15. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL. pp 1746–1751. https://doi.org/10.3115/v1/d14-1181
16. Wang Y, Feng S, Wang D, Zhang Y, Yu G (2016) Context-aware chinese microblog sentiment classification with bidirectional lstm. In: Asia-Pacific Web Conference. Springer. pp 594–606. https://doi.org/10.1007/978-3-319-45814-4_48
17. Wang G, Li C, Wang W, Zhang Y, Shen D, Zhang X, Henao R, Carin L (2018) Joint embedding of words and labels for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. ACL. pp 2321–2331. https://doi.org/10.18653/v1/p18-1216
18. Cui Y, Che W, Liu T, Qin B, Yang Z, Wang S, Hu G (2019) arXiv preprint arXiv:1906.08101
19. Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning. Springer. pp 137–142. https://doi.org/10.1007/bfb0026683
20. Ifrim G, Bakir G, Weikum G (2008) Fast logistic regression for text categorization with variable-length n-grams. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp 354–362. https://doi.org/10.1145/1401890.1401936
21. Johnson R, Zhang T (2017) Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1. ACL. pp 562–570. https://doi.org/10.18653/v1/p17-1052
22. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI Conference on Artificial Intelligence. AAAI. https://ojs.aaai.org/index.php/AAAI/article/view/9513
23. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL. pp 1480–1489. https://doi.org/10.18653/v1/n16-1174
24. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, Wang H (2019) Ernie 2.0: A continual pre-training framework for language understanding. arXiv preprint arXiv:1907.12412
25. Beltagy I, Lo K, Cohan A (2019) Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). ACL. pp 3606–3611. https://doi.org/10.18653/v1/d19-1371
26. Shi H, Xie P, Hu Z, Zhang M, Xing EP (2017) Towards automated icd coding using deep learning. arXiv preprint arXiv:1711.04075
27. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J (2018) Explainable prediction of medical codes from clinical text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. NAACL. pp 1101–1111. https://doi.org/10.18653/v1/n18-1100
28. Zeng M, Li M, Fei Z, Yu Y, Pan Y, Wang J (2019) Automatic icd-9 coding via deep transfer learning. Neurocomputing 324:43–50
29. Li F, Yu H (2020) Icd coding from clinical text using multi-filter residual convolutional neural network. In: Thirty-fourth AAAI Conference on Artificial Intelligence. AAAI. https://doi.org/10.1609/aaai.v34i05.6331
30. Qi L, Wang X, Xu X, Dou W, Li S (2020) Privacy-aware cross-platform service recommendation based on enhanced locality-sensitive hashing. IEEE Trans Netw Sci Eng. https://doi.org/10.1109/TNSE.2020.2969489
31. Wang S, Zhou A, Yang M, Sun L, Hsu C-H, et al (2020) Service composition in cyber-physical-social systems. IEEE Trans Emerg Top Comput 8(1):82–91

32. Wang S, Zhao Y, Xu J, Yuan J, Hsu C-H (2019) Edge server placement in mobile edge computing. J Parallel Distrib Comput 127:160–168
33. Zhou J, Sun J, Cong P, Liu Z, Zhou X, Wei T, Hu S (2019) Security-critical energy-aware task scheduling for heterogeneous real-time mpsocs in iot. IEEE Trans Serv Comput. https://doi.org/10.1109/TSC.2019.2963301
34. Qi L, Hu C, Zhang X, Khosravi MR, Sharma S, Pang S, Wang T (2020) Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment. IEEE Trans Ind Inform. https://doi.org/10.1109/TII.2020.3012157.
35. Zhang S, Li X, Tan Z, Peng T, Wang G (2019) A caching and spatial k-anonymity driven privacy enhancement scheme in continuous location-based services. Futur Gener Comput Syst 94:40–50
36. Zhang S, Wang G, Bhuiyan MZA, Liu Q (2018) A dual privacy preserving scheme in continuous location-based services. IEEE Internet Things J 5(5):4191–4200
37. Zhong W, Yin X, Zhang X, Li S, Dou W, Wang R, Qi L (2020) Multi-dimensional quality-driven service recommendation with privacy-preservation in mobile edge environment. Comput Commun 157:116–123
38. Xu X, Zhang X, Khan M, Dou W, Xue S, Yu S (2020) A balanced virtual machine scheduling method for energy-performance trade-offs in cyber-physical cloud systems. Futur Gener Comput Syst 105:789–799
39. Xu X, Xue Y, Cui M, Yuan Y, Qi L (2019) Joint optimization of energy conservation and migration cost for complex systems in edge computing. Complexity. https://doi.org/10.1155/2019/6180135
40. Zhou J, Wang T, Cong P, Lu P, Wei T, Chen M (2019) Cost and makespan-aware workflow scheduling in hybrid clouds. J Syst Archit 100:101631
41. Xu X, Zhang X, gao H, Xue Y, Qi L, Dou W (2020) Become: Blockchain-enabled computation offloading for iot in mobile edge computing. IEEE Trans Ind Inform 16(6):4187–4195
42. Xu X, Fu S, Qi L, Zhang X, Liu Q, He Q, Li S (2018) An iot-oriented data placement method with privacy preservation in cloud environment. J Netw Comput Appl 124:148–157
43. Chen J, Li K, Rong H, Bilal K, Yang N, Li K (2018) A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Inform Sci 435:124–149

Publisher's Note