International Journal of Hybrid Information Technology
Vol.9, No.4 (2016), pp. 59-68
http://dx.doi.org/10.14257/ijhit.2016.9.4.06

Computer Aided Diagnosis Based on K-means Collaborative Filtering Algorithm

Feng Xue-yuan[2], Li Peng[1,2]* , Qiao Pei-li[2]

[1] *School of Software, Harbin University of Science and Technology, 150080 Harbin, China*
[2] *School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China*
Email:e_roc@126.com

Abstract

In computer aided diagnosis (CAD) process, one of the most challenging problems is data sparsity, which leads to the diagnosis results are not reliable. This paper proposes a clustering collaborative filtering based algorithm to solve the problem of data sparsity. In this paper, we use k-means clustering algorithm to cluster the same type of patients, and then adopt collaborative filtering method to fill the missing data values for each cluster, in this way to reduce the complexity of similarity calculation of collaborative filtering. The proposed method makes full use of the information-sharing mechanism of "similar patient population" to predict and fill the missing values. A hepatitis dataset is used for evaluating the performance of the algorithm. Results indicate that the proposed algorithm has better performance for medical record data sparsity problem.

Keywords: *Collaborative Filtering; Data sparsity; Computer Aided Diagnosis;*

1. Introduction

With the development of IT, the traditional paper-based medical records changed into electronic medical records, so that the computer's information storage capacity, organizational ability and analysis ability have been widely used in the medical field. In particular, the application of data mining techniques in computer aided diagnosis (CAD), which has converted the traditional doctor with experience to determine into a more objective machine judgment. Data mining based CAD (DM-based CAD) is aimed at discovering relevance between diagnosis items and diagnosis results through electronic medical records, and by simple and effective way assists doctors in making reliable decisions, same time tries to avoid because of the subjective factors or incomplete medical records which lead to misdiagnosis[1].

In recent years, many scholars and researchers conducted in-depth exploration and have achieved some progress on improving the performance of CAD. For instance, scholar from Jilin university proposed CAD system based on principle component analysis (PCA) and extreme learning machine (ELM), the algorithms are used in different stages of the diagnostic process, improved the efficiency and performance of the system[2]. Deng combined Co-Forest algorithm and adaptive data editing (ADE) , presented a new co-training-style random forest for CAD called ADE-Co-Forest algorithm to enhance the system learning performance[3]. Spanish scholar Calle-Alonso adopted the Bayesian hybrid classification method in CAD, and the method is based on a hybrid approach which combines pairwise comparison, Bayesian regression and the k-nearest neighbor technique to classify multi-class biomedical objects[4]. Delibasis proposed the of a CAD system based on a supervised classification algorithm with a so-called BoxCells as the core, and the application can improve the accuracy of thyroid

ISSN: 1738-9968 IJHIT

malignancy diagnosis[5]. Then Uetani provided a liver cirrhosis diagnosis method, which converted the segmented liver volume to a triangulated mesh surface and found the principal variation modes, finally to classify the normal and abnormal livers with a mode selection method[6]. Some research results of researchers have been successfully applied to the actual disease diagnosis process and achieved good results, so thank them for the outstanding contributions of medical informatization.

DM-based CAD relies on a large number of electronic medical records, However, with the increase in test indicators, the number of medical record samples required exponential increase, which leads to highly sparse medical record data. Most diagnostic algorithms require sample data is dense in space, then the data sparsity will seriously affect the accuracy of diagnosis. Therefore, a reasonable and effective method of filling missing values can improve the overall quality of the data and improve performance of data mining and diagnosis. In this paper, we conduct in-depth study of data sparsity problem of CAD, then propose a prediction filling method based on clustering and collaborative filtering to fill the missing data values. The method first gather similar patients using k-means algorithm, and then adopt patients based collaborative filtering algorithm to predict and fill the missing data values within each cluster. Ideas of the method is to use the information-sharing mechanism of "similar patient population" to produce more reliable and authentic padding data, in essence solving the medical record data sparsity problem of CAD.

2. CAD Process based on Data Mining

The core idea of DM-based CAD is that: the medical record data as input, through a series of pre-treatment process and excavations to find useful patterns, then with these patterns to imitate doctors to diagnose of thinking and deductive reasoning processes. In addition, the system can keep learning and updating.

2.1 Process Description

DM-based CAD process consists of the following seven phases: diagnosis target recognition, medical record data sampling, data pre-treatment, data patterns mining, diagnosis model creation, results evaluation and decision-making, diagnosis output.

As shown in Figure 1, each stage of the diagnosis process corresponds to a stage of the data mining process: (1)The diagnosis target as a starting point, to select the medical record samples from the target database as operation objects. (2)Sample data pre-treatment, including the elimination of irrelevant attributes (such as phone numbers, etc.), handling of missing values, removing duplicate data, data type conversion,etc. (3)Considering the characteristic of the selected sample data, adopt appropriate data mining methods (classification, clustering, association rules, etc.) and data mining algorithms to generate patterns. (4)According to the patterns or feedback knowledge to establish or adjust the diagnosis model. (5)Evaluate the rationality and accuracy of the results and output stage diagnosis results[7].

Similar to the traditional diagnosis repeated scrutiny and full consideration, the entire CAD process is a dynamic feedback process. Through the evaluation, if the stage results are considered unreasonable , the system needs to repeat the diagnosis process according to the actual situation, until the results are satisfactory and output the final diagnosis results. In addition to the above process, many practical factors should be considered at each stage, as an auxiliary mechanism, which also should be combined with the analysis of the medical staff in the actual process of diagnosis, so as to obtain more reliable results.

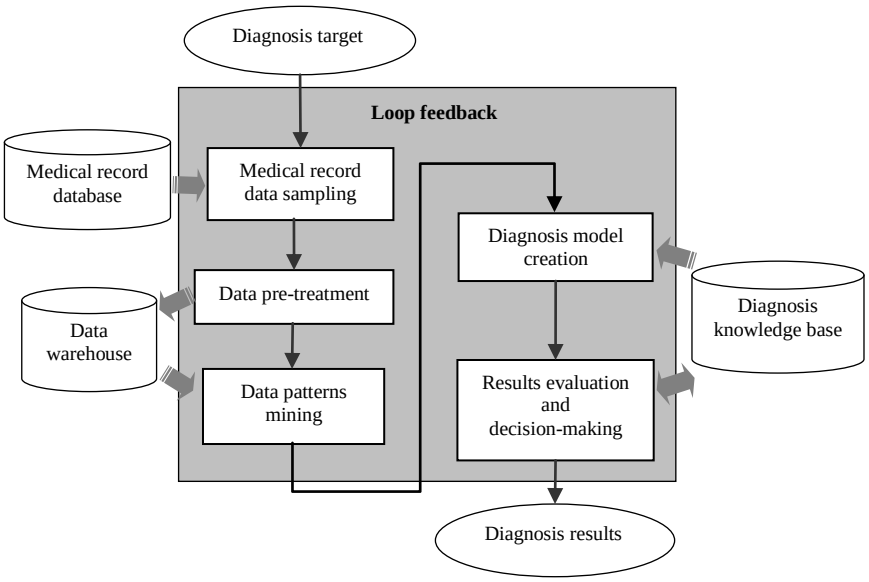

Figure 1. DM-based CAD Process

2.2 Processing of Missing Medical Record Data Values

In the diagnosis process, because not all clinical test results for all patients can be given within a specified time, or there are dependencies between certain items, it will result in some attribute values vacant. So the missing attribute values of medical record data are inevitable, which is the root cause of medical record data sparsity problem. Generally there are several reasons lead to incomplete data, for instance: (1)Temporarily unable to obtain information. (2)Information is missing. (3)One or more attributes of object are not available. (4)Information is ignored. (5)Cost of accessing to information is too big. (6)System real-time requirements are too high[8].

To ensure data integrity, usually missing data values should be properly handled. Common ways to process data missing values are the following: (1)Delete tuple: This method is simple but in some cases will have a great influence on the accuracy of mining results. (2)Filled with special values: The missing values as a kind of special attribute values handling, such as for all missing values use "unknown" to fill, but this may lead to serious data deviation. (3)Mean value method or mode method: They are based on statistics, simple calculations based on other values, but the mean value method will be affected by outliers and easily lead to an unreasonable result, while the mode method under the situation of "no-mode" or "multi-mode" will fail, so such methods have limitations. (4)Decision tree method or regression analysis method: These two methods belong to a prediction filling method, compared with the previous methods, they use multi-directional information of existing data to predict the missing values, but they focus on the estimation of the model parameters instead of missing values, and the modeling process of such methods are too cumbersome.

After in-depth analysis of the advantages and disadvantages of the existing missing value processing methods, we present a new processing method based on clustering collaborative filtering to predict and fill the missing values. The method uses the prediction filling way to process missing values, the idea is that: first use clustering algorithm to gather the same type of patients, and then use collaborative filtering to predict and fill the missing values within each cluster. In this way, the former clustered patients while as possible reducing the latter's calculation. The method is based on the medical record data and uses the similar patients of "collaboration","sharing experiences","giving advice"to fill each missing values. This method abandoned these unreliable elements like "deleting data", "reducing dimension ", "batch unified value",

"subjective filling" in traditional methods. Moreover, the method does not involve complex modeling process. So, we think this is a simple method that can be used to predict and fill the missing values of medical record data, so as to solve CAD data sparsity problem and improve the reliability of diagnosis results.

3. Algorithm Description

3.1. K-means Clustering Algorithm

Clustering is a process that dataset is divided into several groups or clusters, which makes the data objects of the same group with higher similarity and not the same group with lower similarity. There are many clustering methods, such as: k-means algorithm, k-pototypes algorithm, CLARANS algorithm, BIRCH algorithm, CURE algorithm and so on. In which k-means algorithm is the most simple and quick clustering method, and it also has a strong controllability and adaptability[9]. The proposed method just wants to cluster patients roughly , and it requires speediness without the need for accurate clustering results, so k-means algorithm is the best choice.

K-means clustering algorithm is a kind of typical distance-based algorithm, this kind algorithm uses distance as the similarity evaluation index,then the closer the distance of two objects,the higher the similarity of them. Figure 2 shows the use of k-means algorithm to five sample points,in which the number of clusters can be set according to the need.Here k takes 2.

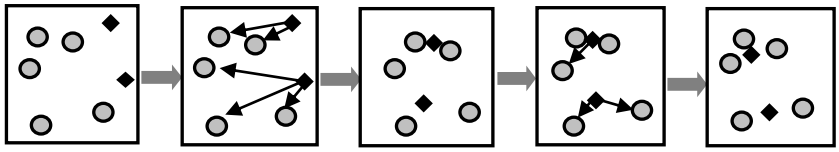

Figure 2. The Process of k-means Clustering

The specific steps are as follows:

1) Random k cluster centroids: $\mu_1, \mu_2, \mu_1, ..., \mu_k \in \mathrm{R}^n$;
2) Repeated until convergence

① For each sample i, calculate its attribution category:

$$c^{(i)} := \arg \min \| x^{(i)} - \mu_j \|^2 \tag{1}$$

② Select all points satisfying the conditions of $c^{(i)} = j$, then get the sum of their values marked as S and the number of these points marked as N;

③ For each cluster j, recalculate the center of the cluster:

$$\mu_j := \frac{S}{N} \tag{2}$$

Among them k is given beforehand the number of clusters; $c^{(i)}$ is the sample i nearest class ($c^{(i)} \in \{1...k\}$); $x^{(i)}$ is the value of i; μ_j is the estimated center of the same kind of samples.

The end condition of k-means algorithm is the result of convergence. As shown in Formula (3), using the distortion function to describe the convergence.

$$J(c, \mu) = \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2 \tag{3}$$

Function J represents the quadratic sum of the distance which is from sample points to their cluster centroid. K-means should adjust J to a minimum. In Formula (3), m is the number of samples.

3.2. Collaborative Filtering Algorithm

Collaborative filtering (CF) is a typical method of using the collective wisdom, which is based on the assumption that: if some objects favor certain items are quite similar, they tend to other items will also be quite similar[10]. For the missing medical record data values, using CF algorithm to predict and fill is based on the assumption: "patients with similar symptoms or indicators tend to show the same trend", which use a similar patient's information-sharing mechanism to cooperate to process the missing values.

CF algorithm, in simple terms is to use the trend of a certain group that has the same favor or a common experience to generate the target object's favor. Meanwhile, the object give considerable response to these items through a cooperation mechanism, then record them in order to achieve the purpose of filtering, thus helping other objects filtering items. CF algorithm is mainly divided into two categories: object-based CF and item-based CF [11].

From the perspective of computing, object-based CF is to take the object's preference for all items as a vector to calculate and obtain its neighbor objects, and then according to the closest neighbors' favor to predict the target object's items that has not been favored. Similarly, the medical records documented a large number of patients' medical history, symptoms, indicators,etc. So first get the most similar patients by the similarity calculation, then according to the "neighbor experience" and by the way of collaboration predict the disease tendency, thus filling the missing medical record data value of the target patient. However, item-based CF is to take all the objects' preference for a certain item as a vector to calculate and obtain the item's neighbor items, and then according to the closest neighbor items and the historic preference of target object to predict. From a medical point of view, there is no necessary connection between of a patient's different test items, or that there is no fixed relative value between items. Therefore, for the prediction filling of the missing medical record data values, object-based CF algorithm is more practical.

There are many similarity calculation methods, we use the Pearson correlation coefficient, as shown in Formula (4), to calculate the similarity between objects. Pearson correlation coefficient is generally used to calculate the linear correlation between two variables of fixed distance, and its value range is [-1, +1]. In terms of the medical record data, Pearson correlation coefficient represents the similarity between the two patient's symptoms, indicators and integrated situations.

The steps of object-based CF algorithm are as follows: (1)Use Formula (4) to calculate the similarity between objects. (2)Select k most similar neighbor objects. (3)Use the nearest neighbor set and Formula (5) to predict the favor of target object.

$$r_{pq} = \frac{\sum_{i \in I_{p,q}} (R_{p,i} - \overline{R_p}) \cdot (R_{q,i} - \overline{R_q})}{\sqrt{\sum_{i \in I_{p,q}} (R_{p,i} - \overline{R_p})^2} \cdot \sqrt{\sum_{i \in I_{p,q}} (R_{q,i} - \overline{R_q})^2}} \tag{4}$$

Here, p and q respectively represent object p and object q ; $I_{p,q}$ represents the intersection of object p and object q of favor items; $R_{p,i}$ represents the favor value of object p to item i ; $R_{q,i}$ represents the favor value of object q to item i ; $\overline{R_p}$ represents the

average favor value of object p to all items; $\overline{R_q}$ represents the average favor value of object q to all items.

$$F_{p,i} = \frac{\sum_{q \in N_p} r_{pq} \cdot R_{q,i}}{\sum_{q \in N_p} | r_{pq} |} \tag{5}$$

Here, $F_{p,i}$ represents the predicted favor value of object p to item i; $R_{q,i}$ represents the favor value of object q to item i, and the q is the neighbor of target object p; N_p represents the neighbor object set of target object p [12].

3.3. K-means CF Algorithm

K-means CF algorithm is based on k-means clustering algorithm and object-based CF algorithm to solve the sparsity problem of medical record data. Idea of the algorithm is that adopt k-means algorithm to cluster patients quickly, then use object-based CF algorithm to predict and fill the missing values of each cluster. In the process of dealing with missing values, the algorithm not only in view of integrity of medical record data and the reliability of padding values, but also in view of the efficiency and adaptability of the algorithm in practical applications.

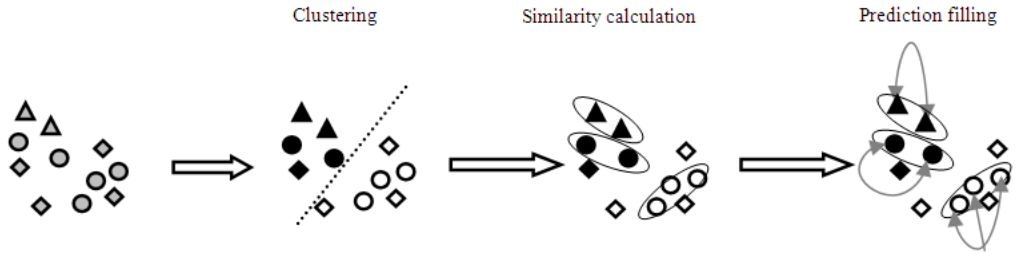

Figure 3. The Process of k-means CF Algorithm

Figure 3 describes the application steps of the k-means CF algorithm in CAD.

Step 1:Cluster patients. The purpose is to roughly clustering the messy medical record data as quickly as possible, so that the patients of the same cluster have high similarity, and patients of different clusters have lower similarity, so the clustering results of this step do not require very precise. In addition, the medical record data is usually constantly changing, while k-means algorithm for dynamic data has a strong ability to adapt, and the number of clusters can be set according to the size of medical record data.

Step 2:Calculate the similarity of patients in the cluster. The algorithm uses Pearson correlation coefficient to measure the similarity of two objects, proceed as follows for each cluster: Apply the Formula (4) for each two objects within the clusters to calculate their similarity, and record their similar values.

Step 3: Predict and fill the missing values. First select a target object, and then follow the k-nearest-neighbor principle to select k most similar neighbors to form the nearest neighbor set. Finally, using Formula (5) and the nearest neighbor set to calculate the missing values of target object and fill them.

After above three steps, that completed the filling of the missing values for one object (target object) of one cluster. Repeating step 3, after each target object within each cluster are filled,this time's prediction filling process are finished. However, in the actual implementation process, just one prediction filling process does not guarantee that all the missing values have been filled, and even a large number of missing values still exist, so need to repeat the the entire prediction filling process. With each time execution of the process, the extent of the data sparse decreases, and the accuracy of clustering and

similarity calculation will become increasingly high, so that the prediction filling results of missing values will be more reliable.

Here, we present the algorithm in detail:

Inputs: Sparse patient-item matrix: $N_{m\times n}$ (the row of the matrix as a vector involved in the calculation), the number of clusters: k, the number of the neighbors to be selected: h $(h<m)$, the sparsity of result matrix: T.

Outputs: The processed matrix: $M_{m\times n}$.

1) Clustering matrix $N_{m\times n}$ into k clusters with k-means algorithm.

2) Perform the following steps for cluster $i(1\le i\le k)$.

① For the cluster i vector $j(1\le j\le n)$, using Formula (4) to calculate the similarity with other $n-j$ vectors, and deposit them into $sim[n][n]$.

② For the matrix sim row l (is also the labeled l vector), select h vectors with larger similarity as the closest neighbors of l, and deposit them into $nei[h]$.

③ Predict the missing values for each vector j of cluster i with Formula (5) and array $nei[h]$, then filling them.

3) Calculate the sparsity of matrix N:

$$S_N=\frac{\text{Number of non-null values of matrix}}{m\times n} \tag{6}$$

4) If $S_N<T$, then put matrix N as an input to re-perform the steps 1) to 3); If $S_N\ge T$, then to perform step 5).

5) Output the final filling results.

4. The Results and Analysis of Experiment

4.1. Dataset

Dataset used in this paper comes from the UCI hepatitis dataset donated by Carnegie-Mellon University researchers. In the hepatitis dataset, a total of 155 cases of hepatitis patients, including 32 cases of died patients. The dataset includes 20 attributes: symptoms, detection index, whether death, and so on. The attribute values of the experimental data have been properly removed, as shown in Table 1, to ensure that dataset with a high sparsity, and Formula (6) is the calculation of sparsity. Meanwhile, in order to easily evaluate the experimental results, the ceiling of the experimental result data sparsity is defined as 0.65.

Table 1. Experimental Dataset

Dataset	Number of Patients	Number of Items	Number of Missing	Sparsity（%）
Hepatitis	155	20	2302	0.2574

4.2. Evaluation Index of Experiment

Here, we use the mean absolute error (MAE) to evaluate the accuracy of missing values filling results, MAE is a measure of the deviation of predicted values with their true specified values. Compared with the average error, MAE is due to changing deviation into

absolute value, and the situation does not occur that the positive and negative offset each other, to some extent, overcomes the drawbacks of the average error. Therefore, MAE for the predicted values can better reflect the actual situation of the error[13].

Let the predicted value set by k-means CF algorithm is $\{x_1, x_2, ..., x_n\}$, and the actual specified value set is $\{y_1, y_2, ..., y_n\}$, then MAE is calculated as follows:

$$MAE = \frac{\sum_{i=1}^{N} | x_i - y_i |}{N} \tag{7}$$

4.3. Experimental Results and Analysis

Experiment respectively used the k-means CF algorithm, mean value method, mode method and regression analysis method for the dataset to predict missing attribute values, and at the same time, ensured that the result matrix of each method has specified and identical sparsity. In order to visualize the experimental results, for the 2302 predicted values generated by the algorithms, using no-repeat simple random sampling method randomly selected 100 points to show the filling results.

Figure 4 shows the MAE values of k-means CF (KCF), mean value method (MVM), mode method (MM) and regression analysis method (RAM), which can compare the effect of each algorithm more precisely.

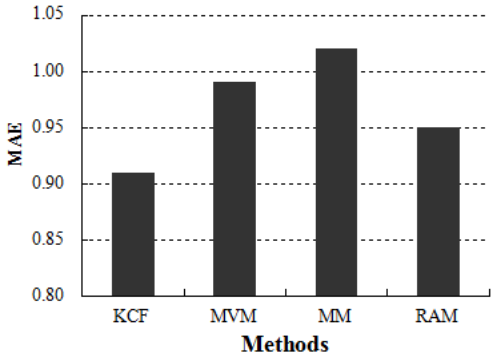

Figure 4. MAE of the Four Methods

As can be seen from the experimental results, the MAE value of k-means CF is the minimum in these four algorithms, which proves compared to the other three methods, k-means CF algorithm indeed can more truly predict missing values of medical record data .Therefore, k-means CF can better solve the data sparsity problem of CAD.

5. Conclusion

At present, the data sparsity problem in CAD has had a great influence on the accuracy of diagnosis results.The paper studied the DM-based CAD process and analyzed the deficiencies of the existing solutions, then proposed k-means CF based algorithm and apply it to CAD for predicting and filling the missing values. By comparative experiment, verified that the predicted result accuracy of the k-means CF algorithm improved 12.1%, laid the foundation for accurate and reliable diagnosis.

However, k-means CF algorithm is sensitive to the sparsity of experimental data, and the experiment also exists the problem of data homologous, so next we will focus on these two factors of the algorithm do further experiment and improvements.

Acknowledgement

This paper is partially supported by Foundation for University Key Teacher of Heilongjiang Province (1252G023). Li Peng is the corresponding author of this paper.

References

[1] A. Osareh, "A Computer Aided Diagnosis System for Breast Cancer", International Journal of Computer Science Issues, vol. 8, no. 2, **(2011)**, pp. 233-240.
[2] L.N. Li, J.H. Ouyang, "A Computer Aided Diagnosis System for Thyroid Disease Using Extreme Learning Machine", Journal of Medical System, vol. 36, no. 5, (2012), pp. 3327-3337.
[3] C. Deng, M.Z. Guo, "A New Co-training-style Random Forest for Computer Aided Diagnosis", Journal of Intelligent Information System, **(2011)**, vol. 36, no. 3, pp. 253-281.
[4] F. Calle-Alonso, "Computer-aided Diagnosis Aystem: A Bayesian Hybrid Classification Method", Computer Methods and Programs in Biomedicine, vol. 112, no. 1, **(2013)**, pp. 104-113.
[5] K. K. Delibasis, P A. Asvestas, "Computer-Aided Diagnosis of Thyroid Malignancy Using an Artificial Immune System Classification Algorithm", IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 5, **(2009)**, pp. 680 - 686.
[6] U. Mei, T. Tomoko, "Statistical Shape Model of The Liver and Its Application to Computer Aided Diagnosis of Liver Cirrhosis", IEEJ Transactions on Electronics Information and Systems, **(2013)**, vol. 133, no. 11, pp. 2037-2043+5.
[7] S. Rahaman, S. Biju, "Data Mining Facilitates E-Patients Through E-Healthcare: An Empirical Study", New Trends in Information and Service Science, NISS '09. International Conference on. IEEE, **(2009)**, pp. 1158 - 1165.
[8] B.-A. David, G. D. Kumar, "Data Envelopment Analysis of Clinics With Sparse Data: Fuzzy Clustering Approach", Computers and Industrial Engineering, vol. 63, no. 1, **(2012)**, pp. 13-21.
[9] X. Zhang, S.Z. Wang, "Research of Text Clustering Based on Fuzzy Granular Computing", Computer Science, vol. 37, no. 2, **(2010)**, pp. 209-211.
[10] W. Suyun, X. Jingjing, "Collaborative Filtering Algorithm Based on Co-clustering Smoothing", Journal of Computer Research and Development, **(2013)**, pp. 163-169.
[11] Z. Yao, Q. Zhang, "Item-based Clustering Collaborative Filtering Algorithm under High-Dimensional Sparse Data", 2009 International Joint Conference on Computational Sciences and Optimization. IEEE, **(2009)**, pp. 787-790.
[12] P. Shi, Z. Zhi-bin, "Collaborative Filtering Algorithm Based on Rating Matrix Pre-filling", Computer Engineering, **(2013)**, vol. 39, no. 1, pp. 175-182.
[13] Q. Luo, X.J. Miao, "Further Research on Collaborative Filtering Algorithm for Sparse Data", Computer Science, vol. 41, no. 6, **(2014)**, pp. 264-268.

