A Content-Boosted Collaborative Filtering Algorithm for Personalized Training in Interpretation of Radiological Imaging

Hongli Lin · Xuedong Yang · Weisheng Wang

Published online: 14 February 2014
© Society for Imaging Informatics in Medicine 2014

Abstract Devising a method that can select cases based on the performance levels of trainees and the characteristics of cases is essential for developing a personalized training program in radiology education. In this paper, we propose a novel hybrid prediction algorithm called content-boosted collaborative filtering (CBCF) to predict the difficulty level of each case for each trainee. The CBCF utilizes a content-based filtering (CBF) method to enhance existing trainee-case ratings data and then provides final predictions through a collaborative filtering (CF) algorithm. The CBCF algorithm incorporates the advantages of both CBF and CF, while not inheriting the disadvantages of either. The CBCF method is compared with the pure CBF and pure CF approaches using three datasets. The experimental data are then evaluated in terms of the MAE metric. Our experimental results show that the CBCF outperforms the pure CBF and CF methods by 13.33 and 12.17 %, respectively, in terms of prediction precision. This also suggests that the CBCF can be used in the development of personalized training systems in radiology education.

Keywords Personalized radiology education · Making error predication · Content-based predictor · Collaborative filtering

H. Lin · W. Wang (✉)
School of Information Science and Engineering, Key Laboratory for
Embedded and Network Computing of Hunan Province, Hunan
University, 410082 Changsha, China
e-mail: wwsh@hnu.edu.cn

H. Lin
e-mail: hllin@hnu.edu.cn

X. Yang
Department of Computer Science, University of Regina, 3737
Wascana Parkway, Regina, Saskatchewan, Canada
e-mail: xuedong.yang@uregina.ca

Introduction

Alleviating variability among radiologists and improving their diagnostic accuracy in the interpretation of radiological imaging have always been a concern in radiology education [1–4]. A number of research results suggest that adequate training is an efficient way to improve the diagnostic performances of radiologists [5, 6]. Now, with the widespread use of the internet in education, computer-aided radiology training approaches are becoming an effective method. The computer-aided radiology training approaches overcome the time, location, and personnel limitations of traditional radiology training [5–7]. Therefore, computer-aided training approaches have become important complements of traditional radiology training.

There exist already a number of services available from the internet and other organizations that provide convenient tools and large numbers of training resources to facilitate training, such as Radiological Association of North America, MedPix™, MyPACS, and SonoWorld. These services allow trainees to benefit from getting access to a wider range of cases, particularly those not typically seen in everyday practice. The training cases are of great value to radiology education. However, maximizing the potential of these resources is still a challenge. Currently, most services available suffer from a common limitation: they, following the one-size-fits-all static training paradigm, present the same content to all trainees regardless of individual needs. Although the services provide tools to enable trainees to retrieve cases that they are interested in, the retrieval tools passively select cases according to the keywords of trainees. In other words, the tools have no capability to recognize the knowledge and experience level of an individual, thus would not be able to intelligently choose suitable cases meeting his or her performance level. We feel that the development of personalized radiology training systems, which intelligently chooses appropriate cases for each

trainee based on individual performance level, deserves more attention from researchers since they make training more efficient in time and effective in outcome.

As for a personalized training system in radiology education, the method of predicting the difficulty level of each case to each trainee will be a core component. Such a prediction algorithm will intelligently evaluate the difficulty level of each case to each trainee and then submits its evaluation results to select optimal cases for each trainee according to his or her performance level. This explains why our work concentrates on the devising predication algorithm for individualized radiology training.

Related Work

Personalized radiology education in the interpretation of radiologic imaging has been gaining increased attention from researchers around the world. Researchers are exploring approaches for the development of personalized radiology education systems for the interpretation of mammography and have made some progresses [8–10]. Sun et al. [8] proposed an ontology to support adaptive training for mammography training. Mazurowski et al. [9] developed a framework for an individualized adaptive computer-aided radiology education system, which utilizes the concept of user modeling to adapt the training protocol to meet the individual needs of the radiologists in training. In another study conducted by Mazurowski and Tourassi [10], they explored the potential of pure collaborative filtering (CF) approach to predict the difficulty levels of unseen cases for a given trainee.

The user model approach belongs to the class of content-based filtering (CBF) algorithms, which uses the error making history of trainees to predict their probabilities of making errors for unseen cases. The fundamental assumption of it is that if two cases are similar, then a trainee will rate them similarly. Therefore, as a content-based prediction algorithm, the user model approach would suffer from the “naive trainee” problem. Stated simply, a new trainee has rated few cases, and hence, his or her trainee-ratings vector is typically very sparse. So, the probability of finding a set of rated cases that is very similar to a target case is lower in the sparse vector. The result will lead to inaccurate predictions, or even failing to predict.

In contrast to the user model method, the CF approach exploits the ratings on a target case, which are given by other trainees, to predict the rating on it that the active trainee will give. The underlying assumption of the CF is that if two trainees are similar in terms of rating cases in the past, then they will be similar in rating new cases as well. The “naive case” issue is a challenge of pure CF algorithms. CF algorithms may fail in making predictions on the new cases due to the lack of ratings on them given by other trainees. In particular, the problem is very significant when the system is in the initial stage of use.

CBF approaches and CF algorithms have both been used fairly successfully to build recommendation systems in various domains [11–13, 16]. However, as described above, they suffer from the cold-start problem in one form or another. This is the well-known problem of handling new cases and new trainees [15]. To overcome the drawbacks of CBF and CF, some hybrid approaches were proposed, which incorporate the advantages of both methods, while not inheriting the disadvantages of either [14, 17–19]. Linear combination models, sequential combination models, and other mixed combination models are three most used strategies. A linear combination model combines the results of a CBF predictor and a CF through an adaptive weighted average to obtain its final prediction, such as in [20]. The sequential combination methods construct first user profiles using a CBF and then apply a CF to make predictions based on those user profiles. Fab [14] is a well-known example of this model. The proposed hybrid methods have been widely used in various domains, especially in e-commerce. However, their potential in radiology education for predicting the difficulty levels of cases for a given trainee has not been exploited so far based on our best knowledge.

In this paper, we propose a content-boosted collaborative filtering (CBCF) algorithm to evaluate the difficulty level of each case for each trainee. Our CBCF algorithm was inspired by a hybrid recommendation method, which was proposed by Prem Melville et al [17]. The new approach utilizes a CBF predictor to convert a sparse trainee-ratings matrix into a dense rating matrix; it then uses a CF approach to provide final predictions. In this way, the naive case and the naive trainee issues are tackled.

The remainder of the paper is organized as follows: Section II describes the two datasets used in our experiment. Section III presents the CBCF algorithm and describes each component of the algorithm in detail. Section IV presents our experimental work including evaluation metrics, experiment design, experimental results, and comparison with pure CBF and CF methods. Section V concludes this paper with a summary of this work along with a discussion of future directions.

Datasets

This work is exempt from human subject’s research regulation. It makes use of a publicly available, completely de-identified lung CT dataset (LIDC) and a teaching and training mammography dataset provided by our cooperating hospital. In the mammography dataset, all cases were completely de-identified.

We developed a radiology training system using the two radiology imaging dataset to train residents or medical students for the interpretation of mammography and lung CT

scans. The mammography dataset includes 200 mammography cases, the lung CT dataset includes 800 lung CT cases, as well as the training process data of 100 trainees (medical students or radiology residents) from the period of April 1, 2012 to May 8, 2012.

The mammography cases were anonymized. The patient name and other private information were replaced with a specific number as the identification, or, simply, null. In each mammography case, the lesion of interest was indicated using a circle. Also, an expert radiologist was required to assign the mammographic breast imaging reporting and data system (BI-RADS) features to assess the likelihood of malignancy of the lesion using a 0–100 probabilistic scale range. The BI-RADS features and their nominal values include mass margin (circumscribed, microlobulated, obscured, indistinct, and spiculated), mass shape (round, oval, lobular, and irregular), mass density (fat-containing, low density, equal density, and high density), and parenchyma density (0–25 %, 25–50 %, 50–75 %, and 75–100 %). Each nominal was assigned an integer value from 1 to 5 to be used in numerical analysis. In our paper, the expert’s assessment was used as the “gold standard” to evaluate the diagnostic correctness of the trainees.

The cases in the lung CT dataset came from the Lung Image Database Consortium (LIDC) [21], a public image database resource for the development of computer-aided diagnosis methods for lung nodule detection, classification, quantitative assessment, and teaching. For each case, four experienced radiologists were asked to outline the boundary of each nodule (the findings) and to assign subjective nodule characteristics with the help of specific annotation software tools. The characteristics and their nominal values include internal structure (soft tissue, fluid, fat, or air), calcification (popcorn, laminated, solid, non-central, central, or absent), sphericity (linear, ovoid, or round), margin (poorly or sharp), spiculation (marked or no spiculation), subtlety (extremely subtle or obvious), and texture (non-solid/ground class opacity, part solid/mixed, or solid texture). Each nominal was assigned an integer value from 1 to 5 to be used in numerical analysis. The radiologists were also required to assess the likelihood of malignancy of each nodule on a 1–5 scale (1: highly unlikely for cancer, 2: moderately unlikely for cancer, 3: indeterminate likelihood for cancer, 4: moderately suspicious for cancer, and 5: highly suspicious for cancer). The assessments of such cases are subjective, and variability exists across radiologists in the task of lung nodule identification, assessment of nodule features, and determining the likelihood of malignancy [23]. In order to reduce the variation between the scores of the four radiologists, only nodules that are marked by all four radiologists as findings were chosen for the training task. The average of the assessment values of all four radiologists was computed, and the normalized average was used as the gold standard of those chosen nodules. The

average method is a commonly used method to get a robust and reliable assessment [9].

The available data also contained the training process data such as the trainees, the cases, the ratings made by trainees, and the time ratings made. In the training process data, the private information of trainees was anonymized. The absolute value of the difference between the assessment of the likelihood of malignancy of a lesion (case) by a trainee and the gold standard was defined as the difficulty level of the lesion for the trainee. For trainee u and case i , the difficulty level of this case for trainee u can be expressed as in Eq. (1), namely

$$R_{u,i} = \frac{|l_{u,i} - l_i|}{|R|}, \quad (1)$$

where $l_{u,i}$ is the assessment of the likelihood of malignancy of case i that trainee u makes, l_i is the gold standard of case i and $|R|$ denotes the rating scale. In this paper, $|R|$ is equal to 1. For example, if a trainee’s assessment of a case is that there is 60 % probability of the lesion being malignant and the gold standard is 75 %, then the difficulty level of this case for this trainee is 15 %. And if another trainee’s assessment to the same case is 70 %, then the difficulty level of this case for him or her is 5 %.

Content-Boosted Collaborative Filtering Algorithm

Our CBCF method combines a CF algorithm and CBF features to improve prediction performance over pure CBF and pure CF algorithms by overcoming the naive trainee and naive case problems. The main idea of the CBCF algorithm is that a pseudo trainee-ratings matrix is constructed through a CBF predictor based on original trainee-ratings data, and then a CF method is used to make a final prediction based on the pseudo training-ratings matrix. In the CBCF method, producing the pseudo trainee-ratings matrix through a CBF predictor and making a final prediction using a CF method are the two core steps of the CBCF. The following describes the implementation of these two steps in detail.

Content-Based Predictor

In the CBCF algorithm, a CBF predictor is used to convert the sparse trainee-case ratings matrix into a dense matrix. The goal of the CBF predictor is to solve the sparsity problem associated with CF algorithms. In this step, a CBF is utilized to produce predictions about the ratings on unseen cases for each trainee. The predictions are then placed in the original trainee-case ratings matrix. Hence, the sparse matrix becomes a dense matrix.

For a target trainee a , our content-based prediction algorithm represents the target trainee's rating as an n -dimensional vector. Each element in it represents the rating on a case given by the target trainee. If the target trainee does not rate a case i , the corresponding element value will be null. To generate the prediction of case i for trainee a , our content-based prediction method looks into the set of cases that the target trainee has rated and computes their similarity to the target case i and then chooses the k most similar cases with the target case i . At the same time, their corresponding similarities are computed.

$$\text{Sim}(i, j) = \cos(F_i, F_j) = \frac{F_i \cdot F_j}{\|F_i\| \|F_j\|} = \frac{\sum_{k=1}^m F_{i_k} \times F_{j_k}}{\sqrt{\sum_{k=1}^m (F_{i_k})^2} \times \sqrt{\sum_{k=1}^m (F_{j_k})^2}}, \quad (2)$$

where “ $\cdot$ ” denotes the dot product of the two vectors and F_i is the feature vector of case i . The lesion appearance descriptors are used as features of each case, and the corresponding integer values of the gold standard are normalized as the feature values. Case i and case j must belong to the same category.

Instead of considering all of the cases as a category, as do in most application domains of content-based filtering, we classified the cases into two categories: the lung CT and the mammography. We argue that it is more reasonable to classify our cases into two types and then to compute the similarity between cases that belong to the same category. In the radiology diagnosis domain, different descriptors are used to assess the lesion's appearance features and to distinguish malignancy from benign. For example, radiologists use the BI-RADS features to describe their assessment of masses in mammography, while completely different descriptors are used to assess the lung nodule in CT scans. Therefore, the method of comparing two cases that belong to the same category should be more reasonable.

Once a set of the most similar cases to the target case are chosen, based on the similarity measure, the next step is to compute the prediction through a weighted sum approach. In the weighted sum approach, the ratings of similar cases are weighted by the corresponding similarity to the target case. Formally, the prediction of case i for trainee u , denoted by $P_{u,i}$, is given by

$$P_{u,i} = \frac{\sum_{l=1}^k (r_{u,l} \cdot \text{Sim}(i, l))}{\sum_{l=1}^k \text{Sim}(i, l)}, \quad (3)$$

Once the most similar cases are found, the prediction is computed by taking a weighted average of the target trainee's ratings on these similar cases.

In this paper, we use the cosine-based similarity computation method to measure the similarity between two cases. In these circumstances, each case is thought of as a vector and the similarity between the two cases is measured by computing the cosine angle between the two vectors of the two cases. Formally, the similarity between case i and j , denoted by $\text{Sim}(i, j)$, is given by Eq.(2), namely

where $P_{u,i}$ is the predication of case i for the trainee u , k is the count of the nearest neighbors of case i , and $r_{u,l}$ is the rating on case l of trainee u . Here, we use a threshold of 0.6 to select the nearest neighbors. For an unrated case in the target trainee vector, if its similar cases are not found in the ratings of the target trainee based on the threshold, then the average rating of the target trainee over the total number of cases will be its prediction value.

Thus, the pseudo trainee-case ratings vector, V_u , consists of the case ratings provided by trainee u , where available, and those predicted by the content-based predictor otherwise, namely

$$V_{u,i} = \begin{cases} r_{u,i} & \text{if trainee } u \text{ rated case } i \\ P_{u,i} & \text{otherwise} \end{cases}. \quad (4)$$

Collaborative Filtering Algorithm

Once the pseudo trainee-ratings matrix is generated, the next step is to make a final prediction through a CF algorithm based on this dense matrix. The CF algorithm first selects the most similar trainees of the target trainee and then computes the final predictions of the unseen cases for the target trainee based on the ratings of his or her nearest neighbors.

In our paper, the Pearson correlation coefficient was used to measure the similarity between two trainees of a given category. Formally, the similarity between trainee u and trainee a of category c , denoted by $P_{u,a,c}$, is given by Eq.(5), namely

$$P_{u,a,c} = \frac{\sum_{i=1}^m (v_{a,i} - \bar{v}_{a,c}) \times (v_{u,i} - \bar{v}_{u,c})}{\sqrt{\sum_{i=1}^m (v_{a,i} - \bar{v}_{a,c})^2} \times \sqrt{\sum_{i=1}^m (v_{u,i} - \bar{v}_{u,c})^2}}, \quad (5)$$

where c is a category of cases, m is the number of cases that belong to category c , $\overline{V_{u,c}}$ is the average rating of category c given by trainee u , and $v_{a,i}$ is the rating on case i in the pseudo trainee-ratings matrix (it is the original rating given by trainee a or a prediction value provided by the CBF predictor).

Instead of computing the similarity between two trainees based on the ratings of all of the cases, we computed the similarity based on the ratings of the cases that belong to the same category. This is a reasonable approach because we argue that correctly assessing different types of radiological images requires different skills. For a trainee, his or her performance levels of diagnosis of different types of radiological images might not be the same. For example, although a trainee is good at diagnosing ultrasound imaging, he or she might not be good at MRI imaging; similarly, a radiologist may have a higher probability of making errors in the interpretation of mammography, but a lower probability in the diagnosis of lung cancer.

Next, a CF algorithm selects the most similar trainees to the target trainee using the nearest neighbor method. Predictions are then computed as the weighted average of the deviations from the neighbor's mean, that is

$$P_{a,i} = \overline{v_{a,c}} + \frac{\sum_{u=1}^k (v_{u,i} - \overline{v_{u,c}}) \times P_{a,u,c}}{\sum_{u=1}^k P_{a,u,c}}, \quad (6)$$

where $P_{a,i}$ is the prediction of case i for trainee a , $P_{a,u,c}$ is the similarity between trainee a and u about category c , and k is the number of trainees in the neighborhood. For our experiment, we use a threshold of 0.6 to select neighbors.

In the CBCF, there are some factors that influence the accuracy of the final prediction. One is the number of overlapping cases between trainees, which are used to evaluate the similarity between trainees. It is recognized that the number of overlapping cases influences the accuracy of prediction. For an active trainee, the neighbors derived from a small number of overlapping cases tend to be poor predictors. To devalue the correlations based on a few co-rated cases, a significance weighting factor can be used. In our experiment, we use $n/30$ as the factor value, where n is the number of the two trainees' overlapping cases. If two trainees have less than 30 co-rated cases, then the factor value is $n/30$. If they have more than 30 co-rated cases, then the factor value is 1. The significance weighting factor can be described as the following:

$$sg_{u,a} = \begin{cases} n/30 & \text{if } n \leq 30 \\ 1 & \text{otherwise} \end{cases}.$$

The second is the trainee-case ratings data. The accuracy of the pseudo trainee-case ratings vector also has an impact on the similarity computation between trainees. If a trainee rated many cases, then the CBF is effective and, hence, his or her pseudo trainee-case ratings vector is fairly accurate. Otherwise, the vector will not be as accurate. Inaccuracies in pseudo trainee-case ratings vectors often yield misleadingly high correlations between the target trainees and other trainees. Hence, the similarity of two trainees was weighted using the harmonic mean weighting factor as follows:

$$hm_{a,u} = \frac{2m_a m_u}{m_a + m_u},$$

$$m_a = \begin{cases} n_a/30 & \text{if } n_a < 30 \\ 1 & \text{otherwise} \end{cases}$$

where n_a is the number of cases rated by the trainee a . The role of the harmonic mean is to bias the weight towards the lower of the two values m_a and m_u .

Therefore, we take into consideration the above described factors that influence the accuracy of final prediction and added these two weight factors to the traditional collaborative filtering prediction. So, the Eq.(6) is modified as follows, that is

$$P_{a,i} = \overline{v_{a,c}} + \frac{\sum_{u=1}^k (v_{u,i} - \overline{v_{u,c}}) \times P_{a,u,c} \times (hm_{a,u} + sg_{a,u})}{\sum_{u=1}^k (P_{a,u,c} \times (hm_{a,u} + sg_{a,u}))}. \quad (7)$$

Experimental Evaluation

Experiment Design

We used a subset of the training process data described earlier as our experimental data. From 100 trainees, we only selected the trainees who rated both the mammography cases and the lung CT cases. In the end, 30 trainees met our selection criteria. Next, 15 trainees were randomly selected as the test trainees from these 30 trainees. The training process data of the test trainees was divided into three test datasets according to the time sequence: dataset A , dataset B , and dataset C . The test datasets are summarized in Table 1.

For each trainee in each test dataset, ratings for 20 % of the cases were withheld. Predictions were computed for the withheld cases in each test dataset using the CBCF algorithm, pure CBF predictor, and pure CF method, separately.

Table 1 Experimental data

	Dataset A	Dataset B	Dataset C
Time span	April 1–April 13,2012	April 1–April 20,2012	April 1–May 8,2012
Number of rated case times	5005	6012	13480
Mammography cases times	2000	2480	3212
Lung CT cases times	3005	3532	10268
Ratings (%)	0–100	0–100	0–100

Evaluation Metrics

There are several measures for evaluating the quality of a prediction algorithm. The mean absolute error (MAE) [22] between ratings and predictions is a widely used metric. MAE measures the deviation of predictions from their true user-specified values. For each ratings-prediction pair $\langle r_i, p_i \rangle$, MAE computes the absolute error between them. This metric is computed by summing these absolute errors of the N corresponding rating-prediction pairs and then computing the average. Formally,

$$\text{MAE} = \frac{\sum_{i=1}^N |r_i - p_i|}{N} \quad (8)$$

The lower the MAE, the more accurately the prediction algorithm predicts the user ratings.

In our paper, we used the MAE as the evaluation metric to report prediction experiments because it is the most used and is easiest to interpret directly.

The Results

Table 2 showcases the statistical accuracy of the different prediction algorithms in terms of the MAE on test datasets A, B, and C respectively. Figure 1 illustrates the MAE values of our CBCF algorithm of each test trainee on each test dataset.

Table 2 shows the MAE of our CBCF, the pure CBF, and the pure CF prediction approaches on test datasets A, B, and C. It reveals that our CBCF algorithm makes, overall, more accurate predictions than the pure CBF and the pure CF on all

Table 2 Statistical accuracy of the different prediction algorithms in terms of MAE

Prediction algorithms	Dataset A	Dataset B	Dataset C
Pure content-based	0.521	0.489	0.488
Pure CF	0.501	0.463	0.401
Content-boosted CF	0.428	0.268	0.304

three test datasets. Specifically, our CBCF performs 9.13 and 7.3 % better than the pure CBF and the pure CF, respectively, on the test dataset A, 22.3 and 19.5 % on the test dataset B, and 9.7 and 18.4 % on the test dataset C.

Figure 1 displays the MAE of the CBCF algorithm of each test trainee on the test dataset A, B, and C. The figure demonstrates that our CBCF method has the capability to boost the prediction precision for these test trainees on each test dataset. These consistent findings suggest that the pure CF, the pure CBF, and the CBCF are all usable to predict the difficulty level of assessing cases for trainees in radiology education. However, the CBCF makes more accurate predictions as compared to the pure CF and CBF.

However, exceptions exist. It is observed in Fig. 1 that all three prediction methods fail at prediction for test trainee 7. In order to understand what causes the unsuccessful prediction for test trainee 7, the numbers of co-rated cases between test trainee 7 and the other test trainees were examined. It was discovered that the number of co-rated cases between test trainee 7 and the other test trainees was very few. As the majority of the prediction algorithms are based on similarity measures computed over the co-rated set of cases, large levels of sparsity of trainee-case ratings can be detrimental to prediction algorithms. We argue that it is the small number of overlapping cases between test trainee 7 and the other test trainees that lead to the poor prediction of test trainee 7.

Please note the interesting fact that the performance of our algorithm does not improve with data size multiplication. The

Fig. 1 The prediction accuracy of CBCF algorithm on test dataset a, b, and c of each test trainee in terms of MAE

MAE of the CBCF increases from 0.268 on test dataset B to 0.304 on test dataset C, instead of falling as the data becomes dense. We think that ignorance of performance improvement of trainees may be a main reason.

Conclusion and Future Work

In this paper, we proposed a CBCF algorithm to predict the difficulty level of each case for each trainee. The CBCF algorithm overcomes the drawbacks of pure CF and pure CBF and incorporates their advantages.

The CBCF approach can overcome the naive case problem of the pure CF algorithm. In pure CF, a prediction cannot be made for a case with respect to the active trainee, unless it was previously rated by other trainees. In CBCF, a CBF predictor can be used to make a prediction on a new case for the active trainee as well as other trainees based on the similarity between previously rated cases and the new case. In this way, the CBCF approach tackles the new case problem.

Also, the CBCF approach solves the naive trainee issue. A naive trainee is one for whom his or her trainee-case ratings vector is sparse. The sparse trainee-case ratings vector leads to failure in finding similar trainees for the naive trainee, because there are very few co-rated cases in the sparse trainee-case ratings matrixes of trainees. Therefore, the sparsity issue, which is caused by naive trainee, is a major challenge in a CF approach. In the CBCF, we use a pseudo trainee-ratings matrix to solve the sparsity problem, since the pseudo ratings matrix is a dense matrix. The pseudo ratings matrix contains ratings for all cases for all trainees; hence, all trainees will be considered as potential neighbors. This increases the chances of finding similar trainees. As the sparsity of the trainee-case ratings matrix affects both the CBCF method and the pure CF algorithm, it affects the CBCF method to a smaller degree than the CF. Thus, the CBCF can produce better predictions than the CF.

Our experimental results showed that the CBCF outperforms the pure CBF predictor and the CF using the MAE metric. This result suggests that the CBCF method can be used in the development of personalized radiology education system for choosing suitable training cases with respect to individual trainees.

When using our CBCF algorithm to generate optimal training plans for a personalized training system, the strategy of recalculating data for active trainees to find better neighbors should be considered. In our CBCF, making predictions for an active trainee is mainly based on the ratings of his or her neighbors, as well the ratings made by him or her. Hence, finding better neighbors for the active trainee is critical. In actual radiology training circumstances, the trainees similar to the active trainee likely change over time as trainees make progress through training at different paces. So, when making

prediction on a target case for an active trainee, the neighbors of the active trainee should be recalculated in order to obtain a more accurate prediction. However, if we were to recalculate the similarity between the active trainee and the other trainees when making each prediction, it will be computationally intensive. Therefore, a more efficient recalculating strategy is required. In our view, a time interval method and a rating data growth method can be considered. In the time interval method, the process of computing the similarity between trainees will be started periodically at a fixed interval. Conversely, in the rating data growth approach, the recalculation program starts as the rating data grows by a certain amount.

The study presented has some limitations. One limitation is the small sample size. In our experiment, the training process data consisted of 15 trainees. In the future, we will focus on collecting more training process data, as well as including more modality of radiology images. The second limitation is ignoring the training performance improvement in our prediction algorithm. In the CBF predictor, the prediction of a target case for an active trainee is based on the previous ratings of similar cases of the target case, which were given by the active trainee at his or her past performance level. Thus, it may affect the prediction accuracy. In the future, we will take into account the performance improvements of trainees to improve the accuracy of our prediction.

Acknowledgments This work was supported by Grant No. 81201151 from Natural Science Foundation of China, No. 12JJ6061 from Natural science foundation of Hunan province, and No. 2012SK3185 from Social development support plan of Hunan province.

References

1. Elmore JG, Wells CK, Lee CH, et al: Variability in radiologists' interpretations of mammograms. *N Engl J Med* 331(22):1493–1499, 1994
2. Watahani T, Sakai F, Johkoh T, et al: Interobserver Variability in the CT Assessment of Honeycombing in the Lungs. *Radiology* 266:936–944, 2013
3. Kerlikowske K, Grady D: Variability and Accuracy in Mammographic Interpretation Using the American College of Radiology Breast Imaging Reporting and Data System. *JNCI J Natl Cancer Inst* 90(23):1801–1809, 1998
4. Jiang Y, Nishikawa RM, Schmidt RA, et al: Potential of Computer-aided Diagnosis to Reduce Variability in Radiologists' Interpretations of Mammograms Depicting Microcalcifications. *Radiology* 220: 787–794, 2001
5. Taylor PM: A Review of Research Into the Development of Radiologic Expertise: Implications for Computer-Based Training. *Acad Radiol* 14(10):1252–1263, 2007
6. Linver MN, Paster SB, Rosenberg RD, et al: Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology* 184:39–43, 1992

7. Leung JWT, Margolin FR, Dee KE, et al: Performance parameters for screening and diagnostic mammography in a community practice: are there differences between specialists and general radiologists. *AJR. Am J Roentgenol* 188:236–241, 2007
8. Sun S, Taylor P, Wilkinson L, and Khoo L: An ontology to support adaptive training for breast radiologists. In *Proceedings of the International Workshop on Digital Mammography*, 2008, pp. 257–264
9. Mazurowski MA, Baker JA, Barnhart HX, Tourassi GD: Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments. *Med Phys* 37(1152):1152–1160, 2010
10. MA Mazurowski, GD Tourassi: Exploring the potential of collaborative filtering for user-adaptive mammography education. *Biomedical Sciences and Engineering Conference (BSEC)*, 2011
11. Allen, RB: User models: theory, method and Practice. *International Journal of Man–Machine Studies*, 1990
12. Herlocker, JL, Konstan, JA, Riedl, J: Explaining collaborative filtering recommendations. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2000
13. Herlocker J, Konstan JA, Terveen L, Riedl J: Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22 (1), 2004
14. Balabanovic M, Sholam Y: Fab:Combining content-based and collaborative recommendation. *Commun ACM* 40(3):66–72, 1997
15. Huang Z, Chen H, Zeng D: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans Inf Syst* 22(1):116–142, 2001
16. Linden G, Smith B, York J: Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7(1):76–80, 2003
17. MelvilleP, Mooney RJ and Nagarajan R: Content boosted collaborative filtering for improved recommendations, in *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI 02)*, Edmonton, Canada, 2002, pp. 187–192
18. Su X, Khoshgoftaar TM, Zhu X and Greiner R: Imputation-boosted collaborative filtering using machine learning classifiers, in *Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC 08)*, Cear'a Fortaleza, Brazil, March 2008, pp. 949–950
19. Pennock, D. M, Horvitz E, Lawrence S and C. L. Giles, Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach, in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI 00)*, 2000, pp. 473–480
20. Claypool M, et al: Combining content-based and collaborative filters in an online newspaper. *Proceedings of ACM SIGIR workshop on recommender systems*. Vol. 60, 1999
21. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, et al: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics* 38: 915, 2011
22. Hofmann, T: Collaborative filtering via gaussian probabilistic latent semantic analysis. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2003