Active Learning for an Efficient Training Strategy of Computer-Aided Diagnosis Systems: Application to Diabetic Retinopathy Screening

C.I. Sánchez[1], M. Niemeijer[2], M.D. Abràmoff[3], B. van Ginneken[1,2]

[1] Department of Radiology, Radboud University Nijmegen Medical Centre, The Netherlands
[2] Image Sciences Institute, University Medical Center Utrecht, The Netherlands
[3] Department of Ophthalmology and Visual Sciences, University of Iowa, USA

Abstract. The performance of computer-aided diagnosis (CAD) systems can be highly influenced by the training strategy. CAD systems are traditionally trained using available labeled data, extracted from a specific data distribution or from public databases. Due to the wide variability of medical data, these databases might not be representative enough when the CAD system is applied to data extracted from a different clinical setting, diminishing the performance or requiring more labeled samples in order to get better data generalization. In this work, we propose the incorporation of an active learning approach in the training phase of CAD systems for reducing the number of required training samples while maximizing the system performance. The benefit of this approach has been evaluated using a specific CAD system for Diabetic Retinopathy screening. The results show that 1) using a training set obtained from a different data source results in a considerable reduction of the CAD performance; and 2) using active learning the selected training set can be reduced from 1000 to 200 samples while maintaining an area under the Receiver Operating Characteristic curve of 0.856.

1 Introduction

In the last decade a variety of computer-aided diagnosis (CAD) systems has been developed for the automatic screening of diverse diseases, such as breast or lung cancer or diabetic retinopathy [1]. In general, these systems receive as input an exam consisting of one or more images from a patient and they generate as output a degree of suspicion for the disease. A typical CAD system relies on multiple stages: segmentation of normal anatomy, localization of abnormalities and fusion of different findings to obtain the final decision.

Supervised classification has been widely adopted for CAD systems as the optimal solution for the fusion and generation of the final outcome. In this type of approach, the training phase is of paramount importance for the development of these systems. During this phase, the classifier 'learns' from a group of *a priori* manually-annotated sample exams, namely training set. This training set should be representative enough to cope with the image variability and the

T. Jiang et al. (Eds.): MICCAI 2010, Part III, LNCS 6363, pp. 603–610, 2010.

range of pathologies usually encountered in medical environments in order to obtain good prediction ability. Therefore, the choice of this training set has a great impact on the final performance. A non-representative training set may dramatically reduce the system accuracy, as we will show in this paper.

The common approach for the selection of training sets in CAD systems is to use public labeled databases or to extract a group of samples from the available data in a random way. This procedure does not always guarantee a representative training set, especially when retrieving data from the large imbalanced databases usually found in CAD applications. This results in the selection of large training sets in order to assure data generalization and high quality results. The annotation of a large amount of data is a tedious and time consuming task for medical experts. Additionally, the computational costs for the training phase increases drastically when the number of training samples increase.

Active learning is a machine learning approach that attempts to retrieve training data maximizing the system performance and minimizing the labeling effort [2]. In active learning approaches, only the most informative samples are dynamically selected from the unlabeled data and their correct labels are requested from an expert. Therefore, the size of the training set required to obtain an optimum classification accuracy is reduced, as well as the user's involvement in the labeling process. Compared to standard classification, where the goal is to minimize classification error, active learning has an additional goal: minimizing the amount of samples to be labeled.

In this paper, we evaluate the benefit of including active learning in the training phase of CAD systems in order to retrieve automatically representative training sets from large medical datasets. Particularly, we assess how the performance of a specific CAD application, namely the screening of diabetic retinopathy, is influenced by the active selection of training samples.

2 CAD System for Diabetic Retinopathy Screening

Diabetic Retinopathy (DR) is the most important cause of blindness in the working population of developed countries [3]. Early detection and diagnosis through screening programs is crucial for the prevention of visual loss and blindness in patients with diabetes [3]. A CAD system for the automatic large-scale screening of DR provides an effective way to obtain an early diagnosis and to prevent future complications. Figure 1(a) shows an example of a retinal image from a diabetic patient with DR.

The proposed CAD system for DR screening relies on four components: quality verification, normal anatomy detection, bright lesion detection and red lesion detection. These components are all based on previous work [4] and are therefore only briefly described here.

Quality verification: This component uses a statistical classifier to obtain the probability that the image quality is sufficient for diagnosis.

Normal anatomy detection: This component identifies blood vessels and the optic disk in retinal images using Gaussian derivative filters and k Nearest Neighbors (kNN) regressor.

Table 1. Set of features for retinal exam classification. PP: Posterior probability

Feature	Description
1	Quality of the exam Q. For exams with more than one image, the quality is given by $\frac{max(Q_i)-min(Q_j)}{2}$, where Q_k is the quality of image k in the exam.
2,3	The sum of all red/bright PPs as a measure for the total lesion load in the exam.
4,5	Highest red/bright PP in the exam.
6,7	Total red/bright lesion load weighted by the size of the detected lesions.
8,9	Average red/bright PP for those lesions with probability higher than 0.
10,11	Standard deviation of the red/bright PPs for those lesions with probability higher than 0.
12-19	A four bin histogram of the PPs of the red/bright lesion candidates.
20-27	A four bin histogram of the lesion area in pixels subdivided by PP.

Red lesion detection: Red lesions are pathological regions that usually appears in the earliest stages of DR. For their detection, a hybrid candidate extractor and a kNN classifier are applied to obtain a likelihood per candidate to be a red lesion.

Bright lesion detection: As well as red lesions, bright lesions are also important signs of DR. The algorithm relies on a candidate extraction step based on pixel classification and a kNN classifier.

The different findings from the aforementioned components are then fused to obtain a final outcome per patient: the likelihood of the patient to be referred to an ophthalmologist due to the presence of DR signs or because of insufficient quality. For the fusion procedure, we calculate a set of features for each exam based on the output from the different components (see Table 1). With these features, a kNN classifier is trained to obtain a probability per exam.

3 Methods

In order to train efficiently the proposed CAD system, an active learning approach is incorporated in the training phase of the fusion strategy. This approach is an iterative procedure where at each iteration the active learner is called to select an unlabeled sample from a pool of unlabeled data and an expert is asked for its label. The idea is to select efficiently a set of training samples from the unlabeled data in an active way to boost the performance of the classifier and reduce the number of samples that need to be labeled.

Assume that a small initial training set X_t, a classifier c, an active or query function F and unlabeled data X_u are given. The query function F assigns a value to each unlabeled sample in X_u depending on how informative the sample is. These values permit ranking the unlabeled objects and selecting the most informative sample x^*, which is expected to improve the classification performance the most [2].

The general framework of the active learning system can be described as follows [2]:

1. Train classifier c on the current training set X_t.
2. Select an object x^* from the unlabeled data X_u according to the active query function F.
3. Ask an expert for the label of x^*. Enlarge the training set X_t and reduce X_u: $X_t = X_t \cup \{x^*\}$, $X_u = X_u \backslash \{x^*\}$.
4. Repeat steps (1)-(3) until a stopping criterion has been reached.

The active function F determines the sampling selection, i.e., decides which sample in the unlabeled data X_u to query next. We investigate two different query functions in this study: uncertainty sampling and query-by-bagging (QBB) sampling.

Uncertainty sampling: This method queries unlabeled samples about which the current classifier is most uncertain and asks for their correct labels [2]. To measure uncertainty, the query function F_{US} can be defined as follows:

$$F_{US} \equiv x^* = argmax_{x_i \in X_u}[-\sum_j (P(w^j|x_i)\log(P(w^j|x_i)]. \quad (1)$$

with $j = 1, ..., c$ and c the number of classes in the classification problem.

QBB sampling: In this sampling technique, a committee of classifiers are trained and the most informative sample is considered to be the sample over which the committee is in most disagreement about how to label [2]. In each round of active learning, X_t is sampling by replacement R times to create R modified training sets $X_t^1, ..., X_t^r, ..., X_t^R$. The classifier c is then trained with each modified set X_t^r to obtain a committee of R classifiers $\mathcal{C} = c^1, ..., c^r, ..., c^R$. To measure the disagreement among committee members, the query function F_{QBB} is defined as follows:

$$F_{QBB} \equiv x^* = argmax_{x_i \in X_u} \frac{1}{R} \sum_{r=1}^{R} \sum_j P(w^j|x_i; c^r) \log \frac{P(w^j|x_i; c^r)}{P(w^j|x_i; \mathcal{C})} \quad (2)$$

with $P(w^j|x_i; \mathcal{C}) = \frac{1}{R} \sum_{r=1}^{R} P(w^j|x_i; c^r)$.

4 Materials and Experiments

Training set: For the creation of the training set, a group of 7500 unlabeled retinal exams (dataset A) were taken from an online retinal screening program [5]. Each exam consists of four images with resolution varying from 768x576 to 2048x1536 pixels. The exams were obtained using multiple types of fundus cameras while the field of view coverage varied between 35 and 45 degrees. A second publicly available dataset (dataset B) of 1200 exams was also used to train the CAD system and compare the performance with the results obtained using active learning [6]. In dataset B, only one image per exam is provided and they are acquired by 3 ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinography with a 45 degree field of view. The images were captured using 8 bits per color plane at 1440*960,

2240*1488 or 2304*1536 pixels. A human expert manually annotated each exam as normal (546 exams) or suspect (654 exams).

Test set: A totally different group of 7500 labeled retinal exams (dataset C) were also taken from the aforementioned screening program [5]. The exams were manually annotated by a human observer as normal (7080 exams) or suspect (420 exams). An exam is considered suspect if any DR signs are present or if the exam is ungradable.

To evaluate the influence of the training set in the CAD performance with and without the active learning approach, several experiments were performed:

Experiment 1: The CAD system was trained using dataset A and evaluated on dataset C. This was an ideal situation where a large labeled dataset from the same distribution was used for training. The area under the Receiver Operating Characteristic (ROC) curve (Az) was used as performance metric.

Experiment 2: The CAD system was trained using dataset B and evaluated on dataset B. This experiment was done to evaluate the robustness of the system on a totally different data distribution. As the training and the test set was the same, a repeated ten-fold cross-validation was performed and the average Az value was used as performance metric.

Experiment 3: The CAD system was trained using dataset B and evaluated on dataset C. This experiment assessed the CAD performance when a different distribution data was used for training. The Az was also used as performance metric.

Experiment 4: The CAD system was trained using a subset retrieved from dataset A using active learning with uncertainty sampling and evaluated on dataset C. For iteration i of the active learner, the query strategy retrieved an exam from dataset A, a human observer annotated the selected exam and the updated training set $X_t^{(i)}$ was used to train the classifier $c^{(i)}$. The classification performance P was evaluated on dataset C based on Az value when the training set reached sizes of $N = 10, 20, ...100, 200, ..., 1000, 2000, ..., 6000$. With these performance values, a learning curve was calculated. The optimal performance was obtained at the point in the learning curve (and at the corresponding training size) for which the performance obtained with the retrieved training set was non-significantly different from the one obtained with the complete dataset A. Due to the number of comparisons (one per each point in the learning curve), the significance level was adjusted by the Bonferroni correction to $p < 0.0021$ $(0.05/24)$. The training strategy was initialized with 10 samples chosen at random from dataset A. To reduce the influence of the random selection, the experiment was run 10 independent times and the results were averaged to obtain a mean Az value as well as the standard deviation per each point in the learning curve.

Experiment 5: The same as experiment 4 but with QBB sampling as the query function.

Experiment 6: The same as experiment 4 but with random sampling as the query function, the approach typically adopted in CAD development. This method randomly selects the next sample from the pool of unlabeled data (dataset A). In this case the experiment was run 100 independent times.

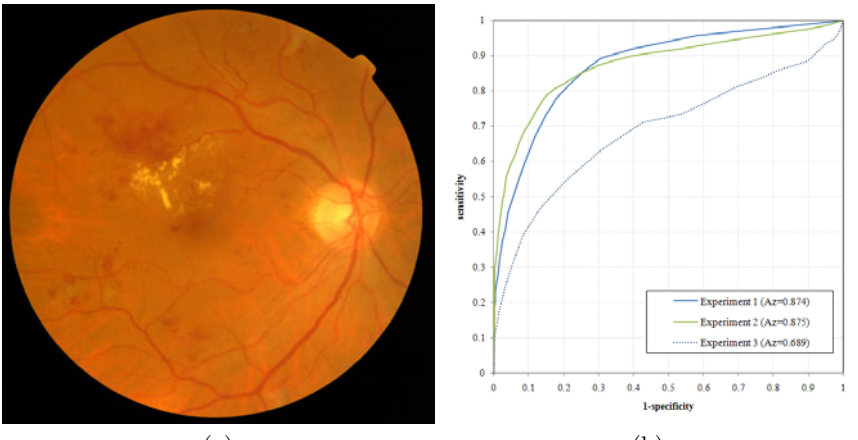
In all the experiments, k was set at the square-root of the number of samples in the training set and, when QBB was used, the number of resampling was set to 3. Previous experiments in a different dataset showed that varying committee size (number of resampling) has little effect on the final performance.

5 Results

Table 2 summarizes the performance of the different experiments. For experiments 4-6, the optimal performance obtained using the learning curves was shown. Figure 1(b) shows the ROC curves for the experiment 1, 2 and 3.

Table 2. Characteristics and performance of the different experiments.* indicates value calculated performing cross-validation.

	Training source	*Training sampling*	*Training size*	*Test set*	*Az*
Experiment 1	Dataset A	-	7500	Dataset C	0.884
Experiment 2	Dataset B	-	1200	Dataset B	0.875*
Experiment 3	Dataset B	-	1200	Dataset C	0.689
Experiment 4	Dataset A	Uncertainty	200	Dataset C	0.856
Experiment 5	Dataset A	QBB	500	Dataset C	0.831
Experiment 6	Dataset A	Random	1000	Dataset C	0.837



(a) (b)

Fig. 1. (a) Example of an image from a diabetic patient with DR. Red and bright lesions appears as red and yellowish patches on the image, respectively. (b) ROC curves for the experiments 1, 2 and 3. In the case of experiment 2, a repeated ten-fold cross-validation was performed.

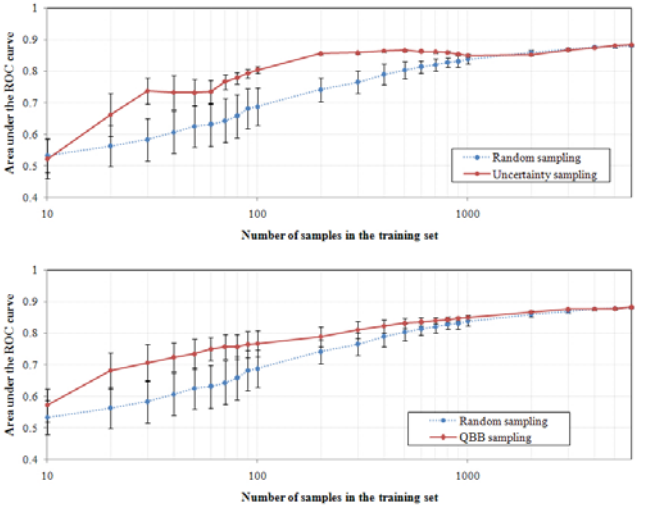


Fig. 2. Learning curves for the proposed active learning strategy using two different sampling functions, namely uncertainty sampling and QBB sampling. The learning curve for a training strategy using random sampling is also depicted. The experiments for each active query function were run 10 times whereas the experiment for the random sampling function was run 100 times. The learning curves show the average area under the ROC curves and the standard deviation of the values obtained for a fixed number of samples in the training set.

Figure 2 shows the learning curves for the experiments 4, 5 and 6. As we can see, at the beginning of the curves the CAD performance was similar to random guess but, after some iteration, CAD performance improved much faster with active learning than using random sampling. The three experiments converged finally to the same point where the training sets were similar to Dataset A.

6 Discussion

In this paper, an effective training strategy for CAD systems using active learning is proposed. Experiment 2 and 3 showed that defining an appropriate training set is a crucial step towards obtaining high quality results in clinical applications. A non-representative training set of the medical data distribution under study had a negative effect on the final CAD performance. In Figure 1, we can see that, although the CAD system worked well for different type of data when the system was trained using samples extracted from the specific distribution, the performance diminished significantly when the training set was from a different distribution. This can probably be attributed to the difference in the number of images, the disease prevalence or the quality between these datasets, which represent normal variations of data characteristics in medical applications. Therefore, this suggested the necessity of creating specific training sets per data type.

Retrieving representative training sets from large unlabeled medical sets is a difficult task, especially when the prevalence of the disease is low. This results in selecting large training set to achieve a good generalization of the distribution, increasing the time and effort spent on performing manual annotations and incurring in higher computational costs for the training phase. We have shown in experiment 4 and 5 that with active learning the training selection rapidly converged to an optimal small training set (see Figure 2 and Table 2). Compared to random sampling, the proposed approach reduced the number of training samples needed to obtain a similar performance when a larger dataset was used, reducing the labeling effort. Although a larger labeled dataset was available for training (dataset B), it was more beneficial to retrieve a small set from the unlabeled set in order to maximize the performance of the CAD system.

However, the selection of the query function can also influence on the final performance. As it is shown in Figure 2, a larger set was needed to obtain the same performance when the query strategy was changed from uncertainty to QBB sampling. Additionally, a stopping criterion needs to be defined for the active learner. Although this criterion can be set depending on a maximum number of manual annotations, this might not guarantee optimal accuracies. In future work, the influence of the stopping criterion will be evaluated.

In conclusion, an active learning approach incorporating in the CAD training phase was studied in order to reduce the number of training samples needed to obtain an optimum accuracy. The results show that the system accuracy can be maximized using small representative training sets, retrieved using an active learner. This approach allows an automatic efficient training stage of the CAD systems using the vast incompletely labeled databases that are now available in medical applications.

References

1. Doi, K.: Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics 31(4), 198–211 (2007)
2. Juszczak, P.: Learning to recognise. PhD thesis, Delft University, the Netherlands, 2006.
3. Kinyoun, J., Barton, F., Fisher, M., Hubbard, L., Aiello, L., Ferris, F.: Detection of diabetic macular edema. Ophthalmoscopy versus photography–Early Treatment Diabetic Retinopathy Study Report Number 5. The ETDRS Research Group. Ophthalmology 96, 746–750 (1989)
4. Niemeijer, M., Abràmoff, M.D., van Ginneken, B.: Information fusion for diabetic retinopathy CAD in digital color fundus photographs. IEEE Transactions on Medical Imaging 28(5), 775–785 (2009)
5. Abràmoff, M.D., Suttorp-Schulten, M.: Web-based screening for diabetic retinopathy in a primary care population: the eyecheck project. Telemedicine Journal and E-health 11(6), 668–674 (2005)
6. Messidor database, `http://messidor.crihan.fr` (accessed March 11, 2010)