Preliminary Investigation on CAD System Update: Effect of Selection of New Cases on Classifier Performance

Chisako Muramatsu*, Kohei Nishimura, Takeshi Hara, Hiroshi Fujita
Department of Intelligent Image Information, Graduate School of Medicine, Gifu University,
1-1 Yanagido, Gifu, Japan 501-1194

ABSTRACT

When a computer-aided diagnosis (CAD) system is used in clinical practice, it is desirable that the system is constantly and automatically updated with new cases obtained for performance improvement. In this study, the effect of different case selection methods for the system updates was investigated. For the simulation, the data for classification of benign and malignant masses on mammograms were used. Six image features were used for training three classifiers: linear discriminant analysis (LDA), support vector machine (SVM), and k-nearest neighbors (kNN). Three datasets, including dataset I for initial training of the classifiers, dataset T for intermediate testing and retraining, and dataset E for evaluating the classifiers, were randomly sampled from the database. As a result of intermediate testing, some cases from dataset T were selected to be added to the previous training set in the classifier updates. In each update, cases were selected using 4 methods: selection of (a) correctly classified samples, (b) incorrectly classified samples, (c) marginally classified samples, and (d) random samples. For comparison, system updates using all samples in dataset T were also evaluated. In general, the average areas under the receiver operating characteristic curves (AUCs) were almost unchanged with method (a), whereas AUCs generally degraded with method (b). The AUCs were improved with method (c) and (d), although use of all available cases generally provided the best or nearly best AUCs. In conclusion, CAD systems may be improved by retraining with new cases accumulated during practice.

Keywords: computer-aided diagnosis system, classifier update, instance selection

1. INTRODUCTION

The idea of computerized medical image analysis has been applied to broad topics with various image modalities and diseases, some of which have already been employed in practice or may be employed in the near future. When a computer-aided diagnosis (CAD) system is implemented in clinical practice, it may be natural that radiologists wish the system would learn from the mistakes. Since clinical cases are continuously accumulated, an effective use of these cases should be considered for improving the system performance and adapting the system to new cases. Therefore, a self-learning CAD system is desired that can be updated automatically and periodically. In general, it is expected that as the number of training cases increases, the computer performs better for new cases, if they are sampled from the same population. However, in some cases, it may be desired to reduce number of training cases for reducing the time of retraining and the workload if user input is required. In such case, new cases to be included in the retraining of a CAD system must be selected so that the retrained system can provide better performance for future cases.

In studies on machine learning algorithms, many groups have investigated instance selection methods for effectively training the algorithms with a reduced number of cases[1-7]. Most of these methods are based on and applied to a k nearest neighbor (kNN) classifier for reducing the case storage space. In several studies, the case selection methods were investigated for a support vector machine (SVM) and other classifiers for reducing the training time and, if possible, excluding noisy data.

In the CAD framework, to our knowledge, not many studies addressed the case selection problem for reducing the storage space, user input, and computational time while maintaining high performance levels. Jamieson *et al* suggested the use of abundant "unlabeled" data, which are without the "gold standard," for system enhancement[8]. By using their method, the system is potentially improved by increasing the number of training cases with reduced labor and time required for data preparation. An "on-line" training technique was suggested by Winter *et al* in which datasets are fed to the system in small batches, and the system is updated whenever a new batch is characteristically different from the already-trained cases[9]. They mentioned that by adopting this method, training is done efficiently, saving computational

*chisa@fjt.info.gifu-u.ac.jp; phone +81-58-230-6519; fax +81-58-230-6514

Medical Imaging 2013: Computer-Aided Diagnosis, edited by Carol L. Novak, Stephen Aylward, Proc. of SPIE
Vol. 8670, 86701T · © 2013 SPIE · CCC code: 1605-7422/13/$18 · doi: 10.1117/12.2007355

time and data storage. Sanchez *et al* investigated the benefit of an "active learning approach" for reducing the number of required training samples with their computerized scheme based on the kNN classifier[10]. They found that the number of training cases could be reduced by "uncertainty" sampling, in which the cases classified near the borderline between positive and negative cases are selected. Mazurowski *et al* investigated different case selection algorithms that are applied to the CAD schemes when using the kNN classifier[11]. Surprisingly, the result suggested that random selection provided comparable or better classification performance than most of the selection methods. One algorithm, i.e., random mutation hill climbing[12], provided comparable classification performance with a large reduction (about 97%) in the number of cases.

In this study, we investigated whether a prototype CAD system can be improved by updating it with new cases as they become available during practice by simulation. As a preliminary investigation, the effect of different case selection methods were compared for investigating the possibility of reducing the number of cases required to update three different classifiers, namely, linear discriminant analysis (LDA), SVM, and kNN. The potential effectiveness of a small step update and a large step update were also examined.

2. MATERIAL AND METHODS

2.1 Classifiers

In this study, the case selection effect was investigated for the task of classifying benign and malignant masses on mammograms. For simplicity, the image features used in the three classifiers were fixed; these features included the effective diameter, degree of circularity, degree of elliptical irregularity, edge contrast, full-width at half-maximum of modified radial gradient histogram, and radial gradient index. The definitions of these features are described elsewhere[13]. Three classifiers from the R programming packages were used in this study. For SVM, a radial basis function kernel was used. For kNN, two tests were performed with a parameter k set to a constant value and a constant fraction of the number of training cases. In this study, k was set to 13 and 6%, respectively, on the basis of the results of five trials in sampling 200 training cases and 100 test cases randomly from the database.

2.2 Database

The database of breast masses used in this study were obtained from the Digital Database for Screening Mammography (DDSM)[14]. For each mass identified in the database, a region of interest (ROI) with a size of 5 cm x 5 cm was obtained. If the outlines of the masses exceeded the size of the ROI, they were not included. Images with architectural distortion and asymmetric density were not used in this study. Other images that were unsuitable for classification, including unclear lesions, lesions with markings, and lesions with a skin fold, were excluded. As a result, the dataset used in this study included 728 and 840 ROIs with malignant and benign masses, respectively.

2.3 Data sampling

For investigating the effect of the case selection methods, a simulation study was conducted by randomly sampling the training and testing datasets from the database. A schematic diagram for the sampling procedures is shown in Fig. 1.

(1) An evaluation dataset E was randomly sampled without replacement in the beginning to keep them independent of the training cases. This dataset was used for evaluating the base classifiers after initial training (Model A) and the updated classifiers (Models B and C).

(2) An initial training dataset I was randomly sampled with replacement from the remaining dataset. This dataset was used for the training of baseline classifiers.

(3) Intermediate testing and training dataset T was randomly sampled with replacement at each update. This dataset was used for testing the previous classifier, and on the basis of this result, some of them were selected (dataset S) to be included in the retraining dataset. In this study, the update (testing and retraining) was performed 10 times. In oppose to the small step updates (Model B), a large step update (Model C), in which all the 10 sampled dataset T's were pooled and used at once, was also investigated.

The above procedures were repeated 100 times to reduce the sampling effect. The numbers of cases in E and T were set to 200 (100 benign and 100 malignant cases), whereas the number of cases in I was varied between 200, 400, and 600. The number of cases in S was varied between 10, 20, and 40. Therefore, the number of cases in T for a large step update

was 2000, and the number of cases in S' was varied between 100, 200, and 400. In this study, the fraction of malignant cases in the sampled datasets was fixed at 50%. The same samples were used for the three classifiers.

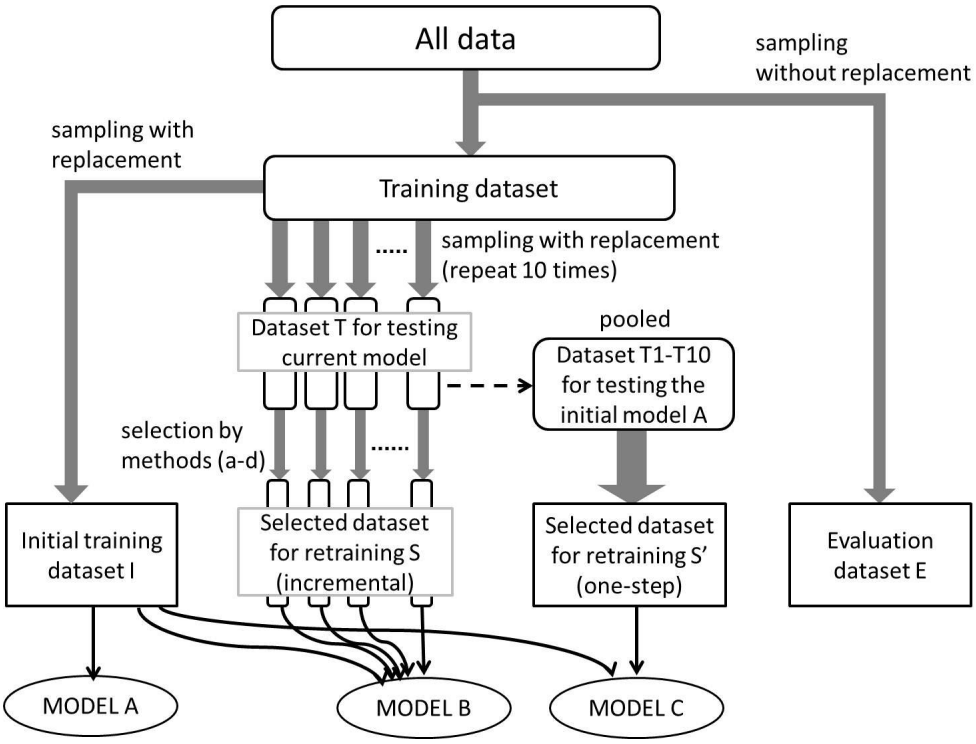

Figure 1. Schematic diagram for sampling procedures.

2.4 Case selection methods

In this study, four case selection methods were considered: (a) cases that were most correctly classified; (b) cases that were most incorrectly classified; (c) cases that were classified as marginal; and (d) random cases.

After each sampling of the dataset T, the current model (model A or models B1-B9) was tested, and the classifiers estimated the likelihood of malignancy for each case. On the basis of these classification results by the current model, the benign and malignant cases with the lowest and highest likelihood of malignancy, respectively, were selected in method (a). Conversely, the benign and malignant cases with the highest and lowest likelihood of malignancy, respectively, were selected in method (b). In method (c), the benign and malignant cases with a likelihood of malignancy around 0.5 were selected.

For comparison, performance by including all cases of T was also determined; however, the number of cases used for training was different (about 3 to 7 times larger) in this scenario. The classifier performance was evaluated by the areas under the receiver operating characteristic curves (AUCs).

3. RESULTS

3.1 Effect of case selection methods

Average AUC after each update by the different case selection methods with LDA, SVM, and kNN when the numbers of cases in I and S were 200 and 20, respectively, is shown in Fig. 2. Note that some results for the method (c) are not shown in the figures because of their low values. The initial AUCs were somewhat lower with kNN, which may be due to the facts that parameter optimization was not sufficient and that the feature set was fixed. In this study, only changes

in AUCs, instead of the absolute values, were considered. In general, the AUCs were almost unchanged with the method (a). With (b), the AUCs slightly improved in the beginning of updates for LDA and SVM; however, the AUCs decreased considerably as the updates were repeated. In the cases of the methods (c) and (d), the AUCs generally improved by updating with all three classifiers, except for kNN with constant *k*. The results suggest that for the classification task used in this study, a small increase in the balanced samples or marginally classified samples has a similar effect by slightly modifying the classification boundary. When using the method (c), the incremental updates may be better than a large step update by adjusting the boundary in small steps instead of drastically changing the boundary. For SVM and kNN with variable *k*, the AUCs by using all cases were significantly higher than those by the selected cases. For LDA, the AUCs by the method (c) were comparable with those by using all cases. The results are summarized in Table 1.

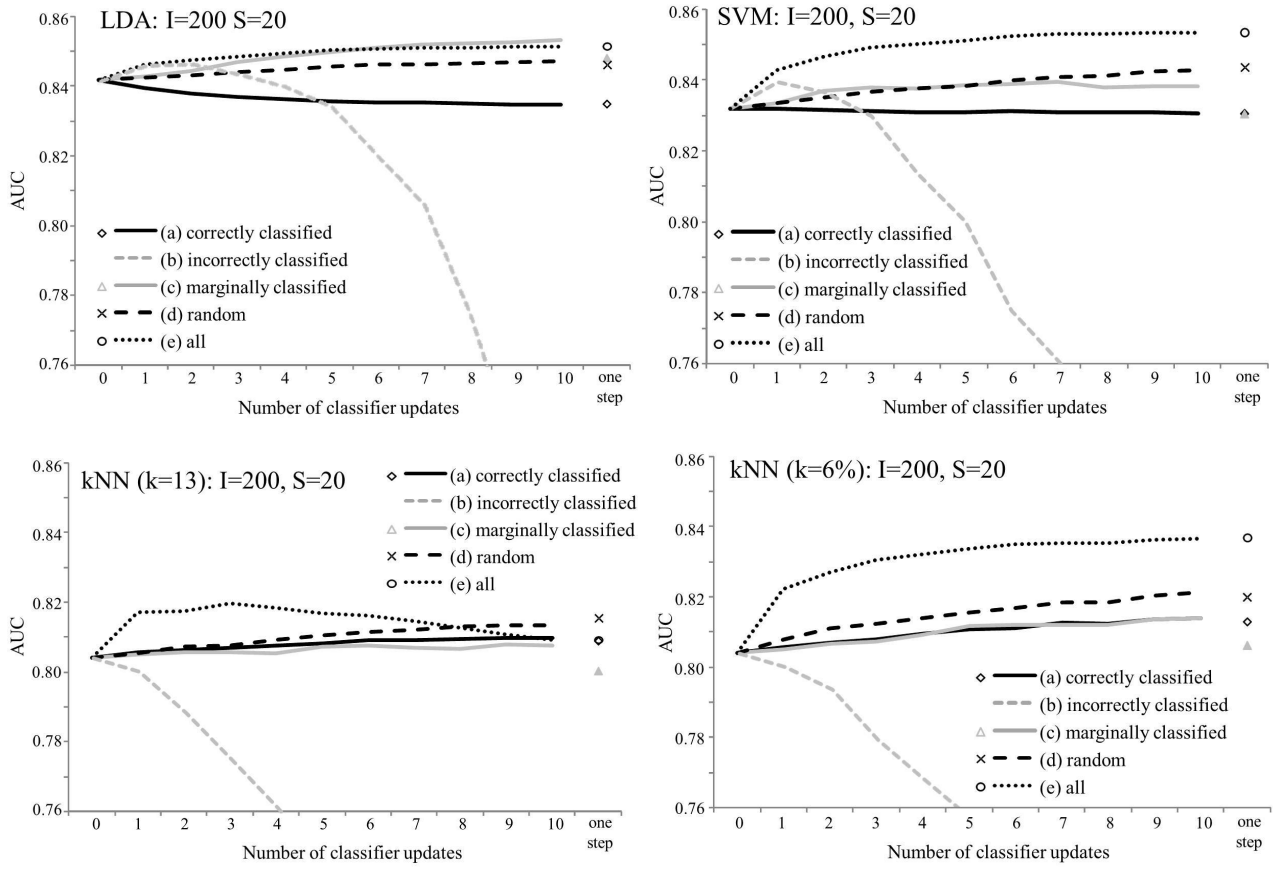

Figure 2. Effect of retraining classifiers on average AUCs by the different case selection methods with dataset I of 200 cases and dataset S of 20 cases (small steps) and 200 cases (one large step).

3.2 Effect of the number of selected cases

When the number of cases used for the updates was varied, a similar trend was observed, however, with the faster improvement or degradation. The results for LDA with dataset I of 200 cases and dataset S of 10 and 40 cases are shown in Fig. 3. It can be observed that the average AUCs by method (b) degrade slower when the number of selected cases was 10 and faster when it was 40 than when it was 20 (See Fig. 2). On the other hand, as the number of selected cases increases, the average AUC by the method (c) approaches and surpasses the result by using all cases at earlier updates.

Table 1. The average and standard deviation of AUCs for initial model A, model B10 (after 10 small-step updates), and model C (one large step update) by the different case selection methods with dataset I of 200 cases and S of 20 cases x 10 updates (small steps) and 200 cases (one step).

Method	Model	LDA	SVM	kNN (k=13)	kNN (k=6%)
	A	0.838±0.03	0.832±0.03	0.800±0.03	0.800±0.03
(a) correctly classified	B10	0.830±0.03	0.829±0.03	0.802±0.03	0.807±0.03
	C	0.830±0.03	0.829±0.03	0.802±0.03	0.807±0.03
(b) incorrectly classified	B10	0.708±0.09	0.731±0.06	0.703±0.04	0.700±0.04
	C	0.455±0.08	0.456±0.07	0.500±0.06	0.474±0.07
(c) marginally classified	B10	0.848±0.03	0.835±0.03	0.807±0.03	0.810±0.03
	C	0.845±0.03	0.827±0.03	0.795±0.03	0.802±0.03
(d) random	B10	0.843±0.03	0.840±0.03	0.810±0.03	0.815±0.03
	C	0.842±0.03	0.836±0.03	0.811±0.03	0.816±0.03
(e) all	B10&C	0.845±0.03	0.848±0.02	0.806±0.03	0.832±0.03

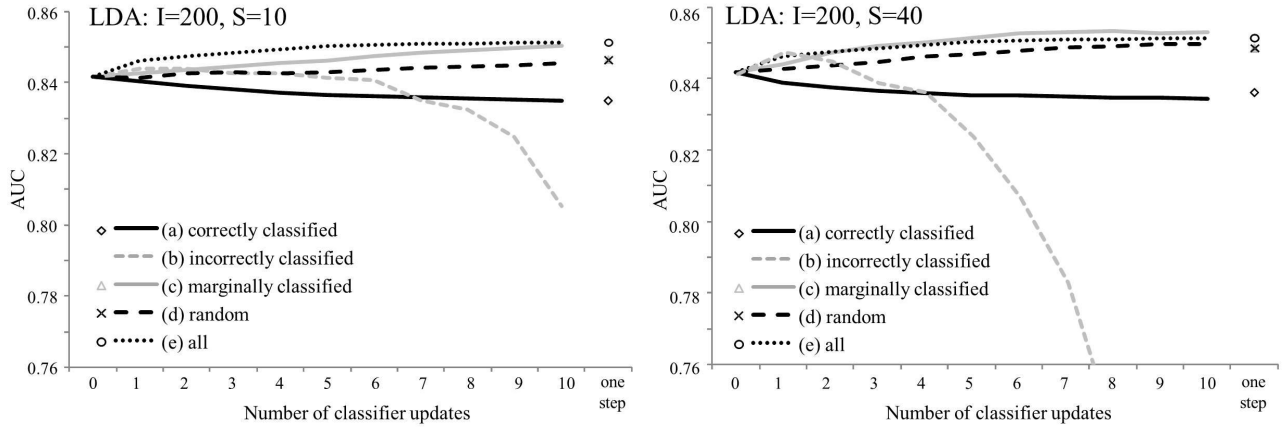

Figure 3. Effect of the number of selected cases of dataset S on average AUCs with dataset I of 200 cases for LDA.

3.3 Effect of the number of initial training cases

As the number of cases in the dataset I was increased, the average AUCs for the initial model A generally became higher with all the classifiers. In addition, the degree of improvement or deterioration in the classification performance due to updates became smaller. The overall trend was similar for SVM; the updated models using all the cases generally provided the highest performance, whereas those using the methods (c) and (d) performed comparably. However, in the case of LDA, the classification performance by the method (b) became considerably higher when the dataset I was large than when it was 200 with the small step updates. Figure 4 shows the results for LDA with dataset I of 400 and 600 cases. It can be predicted that when the fraction of the incorrectly classified cases (atypical or difficult cases) in the training cases is high, the data distribution is altered too much from the population from which test cases are sampled. On the other hand, a small fraction of such cases may provide useful information in adjusting the class border. By the method (c), when the number of cases in dataset I became large, the difference between the AUCs by small step updates and one large step update became small.

In the case of kNN with constant *k*, somewhat different result was observed. When dataset I became large, the AUCs stayed almost constant by the methods (a), (c), and (d); however, the average AUCs obtained by using all cases decreased and became lower than

those by the methods (a), (c), and (d). In fact, this trend was also observed with dataset I of 200 cases (Fig. 2), in which the AUC started to decline as the update proceeded by using all cases. On the other hand, when *k* was varied according with the number of training samples, the expected trend that are similar to those for SVM was observed. Figure 5 shows the results with dataset I of 600 cases and dataset S of 20 cases when the parameter *k* was set of a constant value of 13 and a constant fraction of 6% of the number of training cases. Note that when *k* was constant, the fraction of *k* to the total training samples after 10 updates was 0.5% (13/2600) by using all cases, whereas it was 1.6% (13/800) by selection methods (a) to (d). On the other hand, when *k* was kept 6%, the number of *k* became 49 (800 x 0.06 + 1) and 157 (2600 x 0.06 +1) after 10 updates by the methods (a) to (d) and by using all cases, respectively. These results suggest that the classification performance can be degraded considerably due to a suboptimal parameter of *k*.

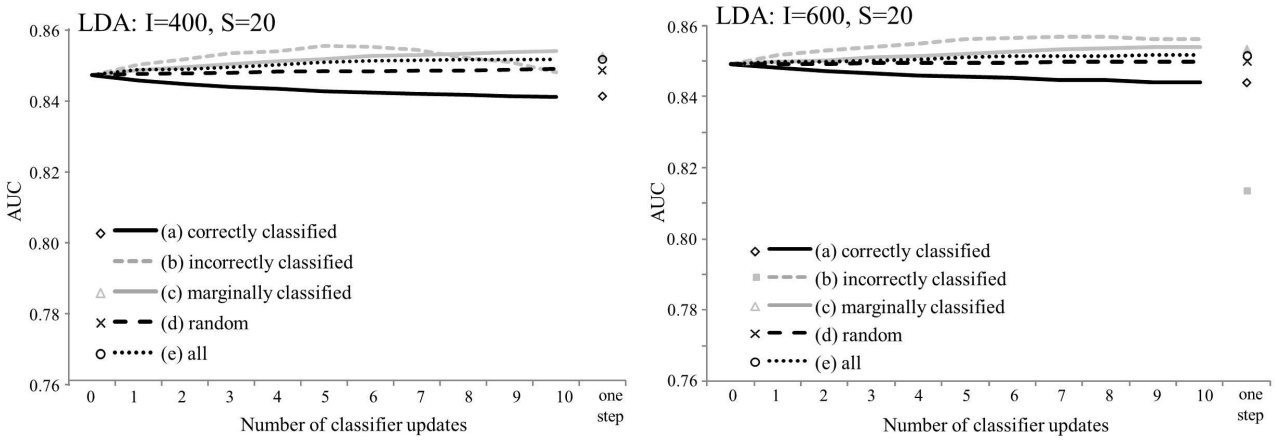

Figure 4. Effect of the number of initial training cases in dataset I on average AUCs with dataset S of 20 cases for LDA.

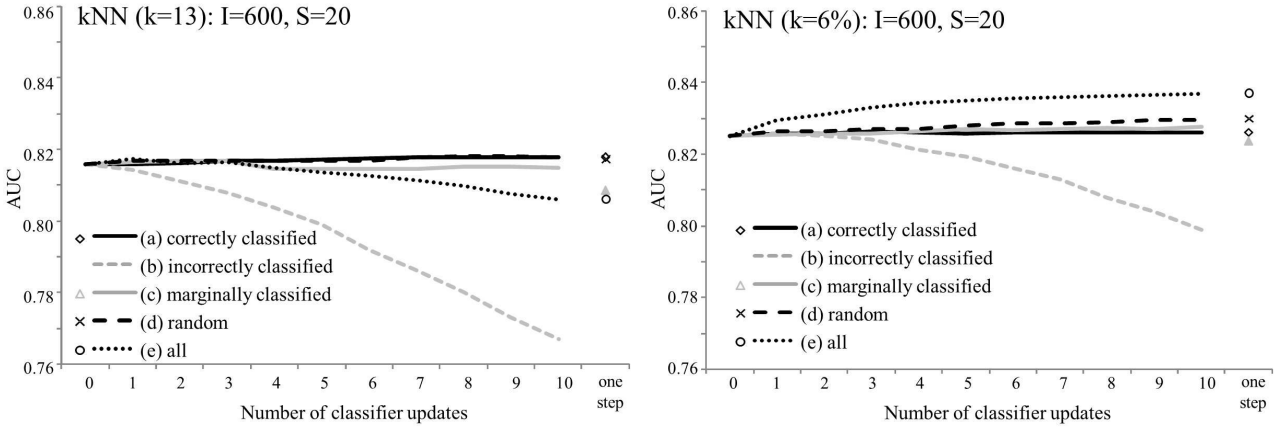

Figure 5. Difference in average AUCs with dataset I of 600 cases and dataset S of 20 cases for kNN with a constant value of k=13 and a constant fraction of k=6%.

4. DISCUSSION

The purpose of this study was to investigate the potential effect of updating CAD systems with newly obtained clinical cases for improving the system. As an initial investigation, the effect of different sampling methods on three different classifiers was studied. Although the use of all the cases would likely result in the best performance, a reduction in the number of cases may be desirable when the preparation of the training cases is time-consuming, and retraining and testing are computationally expensive. In general, the use of correctly classified samples for retraining had small effect. This result was expected because the addition of the correctly classified samples would not alter the classification

boundaries significantly. When using the incorrectly classified samples, detrimental effects may occur if the fraction of these cases in the training dataset is large. The results indicate that once the training dataset is "contaminated," the classification accuracy keeps decreasing. In an attempt to correctly classify a small fraction of extremely difficult cases, a larger faction of typical cases may be classified incorrectly.

The results shown in Fig. 2 and listed in Table 1 indicate that the CAD system can be improved by retraining with the marginally classified cases or randomly selected cases. It was somewhat surprising that the results obtained by using these two methods were comparable. These results may be due to the fact that the number of training cases used in this study was still small. The results were different from those obtained in a previous study[10] in which the use of the marginally classified ("uncertain") cases was effective when compared to the use of the randomly sampled cases. However, the results were somewhat consistent with the results obtained in a study by Mazorowski *et al* in which the results obtained by the use of most of the selection methods were comparable or lower than that obtained by use of the random selection method[11]. In both of these previous studies, the kNN classifier was used. Mazorowski *et al* set the parameter *k* constant, whereas Sanchez *et al* varied the value of k according with the number of training cases. When *k* was kept constant, the results in this study were consistent with those by Mazorowski *et al* until the number of training cases became too large. On the other hand, when *k* was varied, the results of this study were somewhat different from those by Sanchez *et al* in the way that the average AUCs for the randomly selected cases were comparable or higher than for the marginally classified cases. It is not known whether the difference in the results obtained in the previous study and this study was due to the differences in the cases (diseases and image modalities) or the number of cases sampled. The effect of sampling methods must be investigated with different cases in the future.

When the number of cases in the initial training was increased, the effect of retraining was reduced, and the difference between the average AUCs for the different selection methods became smaller. These results could be expected. In the case of LDA, when the fraction of the additional cases was small, the use of the incorrectly classified cases resulted in a beneficial effect. As the number of cases in dataset I was increased, the updated LDA model using the marginal cases started to slightly outperform the model retrained by all the cases. However, it is expected that this effect would be diminished if the number of cases in the initial dataset became too large. In the case of the non-linear classifiers, the classification performances for the models using the marginally classified cases were comparable with those for the models using the randomly selected cases and lower than those for the models using all the cases. Although this study showed some interesting trends and indicated a possibility of improving the classifiers with ideally selected cases, since an optimal fraction of incorrectly classified cases is not known at this time, and the selection of marginally classified cases may require an extra work, use of all cases or randomly classified cases, if case reduction is extremely desirable, is considered the reliable way.

In this study, we examined the effect of using a relatively small number of cases in the initial training and retraining. If the initial training database was very large, there might be almost no effect of retraining the system. However, when it is difficult to collect disease cases or noticeable differences in image characteristics are expected, the incremental system update can be effective. In this study, the selection of the image features and optimization of the parameters were not included in the training. The effect of retraining that includes the search for optimal feature sets and parameters is beyond the scope of this study. The average AUCs for the initial classifiers were somewhat different. The optimizing parameters for each classifier may provide the comparable result; however, in this study, changes in the AUCs, rather than the absolute performance, were considered. Further investigation is certainly needed to draw a strong conclusion.

ACKNOWLEDGMENT

This study was partly supported by a Grant-in-Aid for Scientific Research on Innovative Areas (21103004), MEXT, Japan and a Grant-in-Aid for Scientific Research for Young Scientists (21791179) by Japan Society for the Promotion of Sciences. Authors are grateful to the following radiologists and computer scientist for their contribution in preparation of data: H. Abe, MD, F. Li, MD, and R. Engelmann, MS.

REFERENCES

[1] Hart, P. E., "The condensed nearest neighbor rule," IEEE Trans. Inf. Theory 14, 515-516 (1968).

[2] Wilson, D. L., "Asymptotic properties of nearest neighbor rules using edited data," IEEE Trans. Syst. Man Cybern. 2, 408-421 (1972).

[3] Wilson, R., Martinez, T. R, "Reduction techniques for instance-based learning algorithms," Mach. Learn. 38, 257-286 (2001).

[4] Ferrandiz, S., Boulle, M., "Bayesian instance selection for the nearest neighbor rule," Mach. Learn. 81, 229-256 (2010).

[5] Nikolaidis, K., Goulermas, J. Y., Wu, Q. H., "A class boundary preserving algorithm for data condensation," Pattern Recognition 44, 704-715 (2011).

[6] Mundra, P. A., Rajapakse, J. C., "Gene and sample selection for cancer classification with support vectors based t-statistic," Neurocomputing 73, 2353-2362 (2010).

[7] Li, Y. H., Maguire, L., "Selecting critical patterns based on local geometrical and statistical information," IEEE Trans. Pattern Anal. Mach. Intell. 33, 1189-1201 (2011).

[8] Jamieson, A. R., Giger, M. L., Drukker, K., Pesce, L. L., "Enhancement of breast CADx with unlabeled data," Med. Phys. 37, 4155-4172 (2010)

[9] Winter, L., Motai, Y., Docef, A., "On-line versus off-line accelerated kernel feature analysis: application to computer-aided detection of polyps in CT colonography," Signal Processing 90, 2456-2467 (2010).

[10] Sanchez, C. I., Niemeijer, M., Abramoff, M. D., van Ginneken, B., "Acive learning for an efficient training strategy of computer-aided diagnosis systems: Application to diabetic retinopathy screening," MICCAI LNCS 6363, 603-610 (2010).

[11] Mazorowski, M. A., Malof, J. M., Tourassi, G. D., "Comparative analysis of instance selection algorithms for instance-based classifiers in the context of medical decision support," Phys. Med. Bio. 56, 473-489 (2011).

[12] Skalak, K. B., "Prototype and feature selection by sampling and random mutation hill climbing algorithms," Proc. Int. Conf. Mach. Learn. 1994, 293-301 (1994).

[13] Muramatsu, C., Li, Q., Suzuki, K., Schmidt, R. A., Shiraishi, J., Newstead, G. M., Doi, K., "Investigation of psychophysical similarity measure for evaluation of similar images for mammographic masses: preliminary results," Med. Phys. 32, 2295-2304 (2005).

[14] Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, Jr. P., "The digital database for screening mammography," Proc. International Workshop on Digital Mammography, 2001, 212-218 (2001).