IOP Publishing Phys. Med. Biol. 63 (2018) 115003 (17pp) https://doi.org/10.1088/1361-6560/aac042

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
29 January 2018

REVISED
6 April 2018

ACCEPTED FOR PUBLICATION
26 April 2018

PUBLISHED
22 May 2018

PAPER

Minimization of annotation work: diagnosis of mammographic masses via active learning

Yu Zhao[1,4], Jingyang Zhang[1], Hongzhi Xie[2,3], Shuyang Zhang[2] and Lixu Gu[1,3]

[1] Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China
[2] Department of Cardiology, Peking Union Medical College Hospital, Peking 100005, People's Republic of China
[3] Authors to whom any correspondence should be addressed.
[4] Author contributed to this work.

E-mail: lereinion@163.com (Y Zhao), J.Y.Zhang@sjtu.edu.cn (J Zhang), xiehongzhi@medmail.com.cn (H Xie), shuyangzhang80@gmail.com (S Zhang) and gulixu@sjtu.edu.cn (L Gu)

Keywords: mammographic diagnosis, two-view, ordinal regression, active learning, ranking aggregation

Supplementary material for this article is available online

Abstract

The prerequisite for establishing an effective prediction system for mammographic diagnosis is the annotation of each mammographic image. The manual annotation work is time-consuming and laborious, which becomes a great hindrance for researchers. In this article, we propose a novel active learning algorithm that can adequately address this problem, leading to the minimization of the labeling costs on the premise of guaranteed performance.

Our proposed method is different from the existing active learning methods designed for the general problem as it is specifically designed for mammographic images. Through its modified discriminant functions and improved sample query criteria, the proposed method can fully utilize the pairing of mammographic images and select the most valuable images from both the mediolateral and craniocaudal views. Moreover, in order to extend active learning to the ordinal regression problem, which has no precedent in existing studies, but is essential for mammographic diagnosis (mammographic diagnosis is not only a classification task, but also an ordinal regression task for predicting an ordinal variable, viz. the malignancy risk of lesions), multiple sample query criteria need to be taken into consideration simultaneously. We formulate it as a criteria integration problem and further present an algorithm based on self-adaptive weighted rank aggregation to achieve a good solution.

The efficacy of the proposed method was demonstrated on thousands of mammographic images from the digital database for screening mammography. The labeling costs of obtaining optimal performance in the classification and ordinal regression task respectively fell to 33.8 and 19.8 percent of their original costs. The proposed method also generated 1228 wins, 369 ties and 47 losses for the classification task, and 1933 wins, 258 ties and 185 losses for the ordinal regression task compared to the other state-of-the-art active learning algorithms.

By taking the particularities of mammographic images, the proposed AL method can indeed reduce the manual annotation work to a great extent without sacrificing the performance of the prediction system for mammographic diagnosis.

1. Introduction

Breast cancer is becoming an increasingly severe health problem worldwide (Oliver *et al* 2010). Clinical data show that only early diagnosis and treatment can reduce its mortality rate (International Agency for Research on Cancer 2012). Many technologies that integrate mammography and computer-aided diagnosis have been developed, and are designed to better support doctors and automatically obtain early diagnostic results with greater accuracy (Irwig *et al* 2004, Malich *et al* 2006, Chhatwal *et al* 2009). Supervised learning models such as

support vector machines (SVM) have been extensively employed in diagnosis technologies for mammography and are able to bridge the semantic gap between mammograms and their information regarding diagnosis (Gayathri *et al* 2013). In general, the prerequisite for training an accurate supervised learning model is a large amount of labeled data. Therefore, this requires abundant samples of mammographic images with their corresponding labels provided by radiologists. However, in practical settings, first-hand mammographic images are often unlabeled, and the extensive annotation work required is expensive, time-consuming and requires a vast amount of specialized knowledge.

Recent progress in active learning (AL) algorithms can alleviate this problem by minimizing the labeling costs. As a semisupervised learning method, AL can intelligently choose small 'valuable' subsets from the entire dataset using a specific sample query criterion (SQC) and thus may potentially be used to develop accurate prediction models with less labeling by domain experts required (Panda *et al* 2006, Settles 2012). However, for the problem of mammographic diagnosis, the existing AL algorithms designed for general problems cannot be directly utilized because mammographic diagnosis has two particularities:

(1) Existing AL algorithms assume that the unlabeled samples at hand are independent. For the mammographic diagnosis problem, this assumption may not be valid. There is a definite correlation between two mammographic images because these two images may be of the same suspicious lesion from different views, e.g. the mediolateral view (MLO) and the craniocaudal view (CC), which share the same label and common structural information. A previous study (Gupta *et al* 2006) of supervised learning in mammographic diagnosis indicated that their classification model incorporating both views performed marginally better than the model including only one view. Therefore, we also believe that the results of AL algorithms may be further enhanced if the two-view of mammographic imaging is considered.

(2) Mammographic diagnosis may be regarded as a binary classification task if the user only focuses on the pathological type of each image (mass or normal tissue). However, for the mammographic diagnosis, the researchers would also like to obtain a learning model that can estimate the malignancy risk of breast cancer. This article suggests that this task can be defined as an ordinal regression task rather than a multiclass classification task or an ordinary regression task, because the histological grade is defined on an arbitrary scale in which only the relative ordering between different values is significant. To the best of our knowledge, no related study has combined ordinal regression and AL algorithms. Although there are some similarities between ordinal regression and multiclass classification, the SQC specifically designed for the latter is not appropriate for the former. The reasons for its inapplicability are explained in the following sections. In short, this paper suggests that the selected samples for the ordinal regression model must meet at least two SQCs simultaneously, viz., uncertainty and diversity.

Focusing on these issues, we propose a novel AL algorithm specifically designed for mammographic images, termed 'mammographic diagnosis-based active learning (MDAL)', which is the major contribution of this thesis. MDAL can fully utilize the combining of information from multiple mammographic views by its redefined discriminant functions and modified SQC. Meanwhile, to be able to extend the AL method algorithms to not only classification but also the ordinal regression task, MDAL must be designed to contain more than one SQC, and we formulated the integration criteria strategy (ICS) in the process of MDAL as a rank aggregation problem with self-adaptive weighting and introduced an improved Markov chain to solve this problem, which guaranteed that the selection of mammographic images in MDAL were both representative and informative.

Several comparative experiments on a large database of mammographic images from DDSM, with 900 histograms of oriented gradients (HOG) (Chris Rose *et al* 2006),were also conducted to demonstrate that users can establish a better prediction model for mammographic diagnosis with far less annotation work by employing MDAL. Moreover, this specially designed MDAL has outperformed state-of-the-art AL methods in both classification and ordinal regression tasks for mammographic diagnosis.

2. Related work

2.1. Active learning

The process of AL algorithms is described in figure 1 (see Settles (2012) for more information). In each iteration, the AL algorithm selected the most valuable samples from an unlabeled dataset to query labels using its corresponding SQC. The newly labeled samples were added to a labeled dataset and updated the model of the SQC for the next iteration. According to the corresponding SQC with different definitions of 'valuable', the existing AL algorithms were divided into three categories: representativeness, informativeness and ambiguity measure-based. The AL algorithms in the first category relied on the native data structure, and the samples that represented the majority of the samples were regarded as the most representative, e.g. TED (Yu *et al* 2006b),

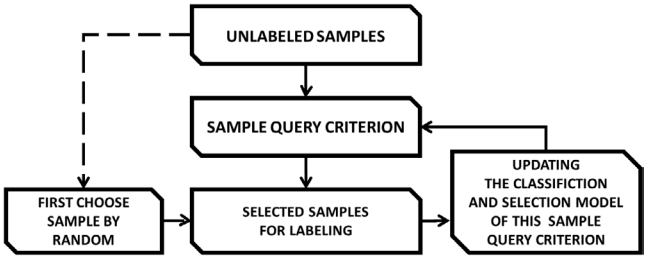

Figure 1. Procedure of the AL algorithm.

MAED (Cai and He 2012), cluster (Dasgupta and Hsu 2008b), density (Jiang and Qing-Yu 2015) and diversity (Demir *et al* 2011). The AL algorithms in the second category always selected the samples that were the closest to the decision boundary or able to impart the greatest change to the current mode, e.g. margin sampling (Lewis and Catlett 1994), estimated error reduction (Roy and McCallum 2001), entropy and expected gradient length (Settles *et al* 2008). The final category selected high-quality samples based on the controversy of multiple learning models, e.g. multiple-view (Muslea *et al* 2006) and query by committee (Freund *et al* 1997). In addition, recent studies have developed a new form of AL algorithm, termed multiple query criteria active learning (MQCAL). By contrast to conventional AL methods, the MQCAL employs more than one SQC, and only the samples that simultaneously meet all of the criteria are selected for labeling (Huang *et al* 2010). According to the different ICSs used for combining all of the complementary information for each involved SQC, typical MQCAL algorithms included MCBAL (Shen *et al* 2004), DUAL (Donmez *et al* 2007) and QUIRE (Huang *et al* 2010). Our proposed MDAL can be regarded as an ad hoc MQCAL with special ICS.

2.2. AL for ordinal regression

All papers cited above were specific to the binary classification problem. However, the practical demands of AL algorithms are not limited to the above scenarios, such as AL for regression and AL for multiclass classification. Three studies implemented AL algorithms in regression problems: linear regression (O'Neill 2015), kernel ridge regression (Douak *et al* 2013), and logistic regression (Schein and Ungar 2007). Three additional studies proposed a combination of AL algorithms and multiclass classification (Joshi *et al* 2009, Demir *et al* 2011, Guo and Wang 2015). To the best of our knowledge, no related study has yet combined ordinal regression and AL algorithms.

As a variation of the one-versus-rest (OVR)-based multiclass classification with one extra order constraint, ordinal regression is used to predict the behavior of ordinal level-dependent variables (predicted labels) with a set of independent variables. In other words, the predicted labels in ordinal regression not only represent the different types of independent variables as multiclass classification but also their ranking or order (Winship and Mare 1984). As for the mammographic diagnosis problem, its predicted label is the degree of malignancy risk of mammary tissue, and thus the application of ordinal regression in mammographic diagnosis problems is more persuasive than multiclass classification.

Assume a learning problem that involves n samples x_i ($i \in [1,2,\ldots, n]$) with their possible r categories ($[c_1, c_2, \ldots, c_r]$). The ordinal regression and multiclass classification model based on logistic regression are then determined by solving the following formulas (1) and (2), more detailed information about the abbreviation and notations in this paper can be found in the appendix.

$$\begin{cases} \ln\left(\frac{P(y\leqslant c_1)}{P(y>c_1)}\right) = \ln\left(\frac{\pi_1}{\pi_2+\cdots+\pi_r}\right) = b_1 + w_1X_1 + w_2X_2 + \cdots + w_pX_p \\ \ln\left(\frac{P(y\leqslant c_2)}{P(y>c_2)}\right) = \ln\left(\frac{\pi_1+\pi_2}{\pi_3+\cdots+\pi_r}\right) = b_2 + w_1X_1 + w_2X_2 + \cdots + w_pX_p \\ \cdots \\ \ln\left(\frac{P(y\leqslant c_{r-1})}{P(y>c_{r-1})}\right) = \ln\left(\frac{\pi_1+\pi_2+\cdots+\pi_{r-1}}{\pi_r}\right) = b_{r-1} + w_1X_1 + w_2X_2 + \cdots + w_pX_p \end{cases} \quad (1)$$

$$\begin{cases} \ln\left(\frac{\pi_1}{1-\pi_1}\right) = b_1 + w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p \\ \ln\left(\frac{\pi_2}{1-\pi_2}\right) = b_2 + w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p \\ \cdots \\ \ln\left(\frac{\pi_r}{1-\pi_r}\right) = b_{(r-1)} + w_{(r-1)1}X_1 + w_{(r-1)2}X_2 + \cdots + w_{(r-1)\,p}X_p \end{cases} \quad (2)$$

where $\boldsymbol{X} = \varphi(\mathrm{x}) = [X_1, X_2, \ldots, X_p]$ denotes the p dimensional feature vector of $\boldsymbol{x}$ in one type of feature space and π_i is the conditional probability $\pi_j = \mathbf{P}(y = c_j|\boldsymbol{X})$. $\boldsymbol{w}$, $\boldsymbol{b}$ are the weights of the prediction model that need to be

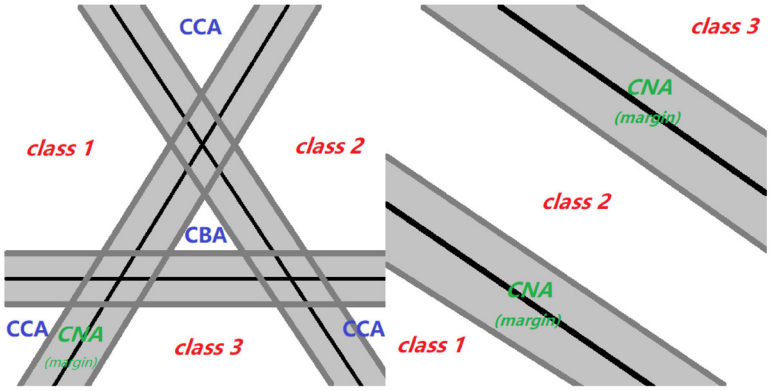

Figure 2. The different discrete sample spaces with hyperplanes based on multiclass classification (left) and ordinal regression (right).

calculated, where $\boldsymbol{b} = [b_1, b_2, \ldots, b_{(r-1)}]$, $\boldsymbol{w} = [w_1, w_2, \ldots, w_p]$ for the ordinal regression model and $\boldsymbol{w} = [w_{11}, w_{12}, \ldots, w_{1p}; w_{11}, w_{22}, \ldots, w_{2p}; \ldots; w_{(r-1)1}, w_{(r-1)2}, \ldots, w_{(r-1)p};]$ for the multiclass classification model [29, 31].

Although formulas (1) and (2) are similar in structure, the different requirements of $\boldsymbol{w}$ render their hyperplane distributions (the black lines) in the discrete sample space entirely different, as presented in figure 2 below. Based on the literature (Guo and Wang 2015), this paper suggests that the selection criterion of the AL algorithm for the multiclass classification problem should select the samples located at the 'disputed areas', which have samples with a high degree of uncertainty and could be considered valuable. These disputed areas include the intersection of classification uncertain areas (CNAs) either with classification blind areas (CBAs) or with region classification compatible areas (CCAs). With regard to the ordinal regression problem, owing to the same $\boldsymbol{w}$ but different b_j, the hyperplanes in its sample space are parallel to one another, which indicates that only several CNAs and neither CCAs nor CBAs exist. Therefore, the AL algorithm designed for multiclass classification based on these three areas is unusable in the ordinal regression task.

Inspired by the literature (Guo and Wang 2015), this article suggests that an appropriate AL algorithm for the ordinal regression should have two simultaneous sample query criteria—uncertainty and diversity. The former criterion guarantees that the selected samples will be located as close to the hyperplanes as possible, and the other ensures a certain number of selected samples around each hyperplane rather than samples around only one hyperplane. Consequently, the AL algorithm for ordinal regression is an MQCAL problem. The uncertainty and diversity-based sample query criteria should be efficiently combined by one ICS, and only the samples that meet both the diversity and uncertainty criteria are selected for querying labels.

2.3. AL in mammographic diagnosis

Few research studies have considered the perspective of applications of AL algorithms in medical tasks, not to mention mammographic diagnosis. In a recent article (Zhou *et al* 2017), AL was introduced for use in convolutional neural networks (CNN) for biomedical image analysis and was demonstrated to perform well in three different biomedical imaging applications. Another work (Zhu *et al* 2014) introduced a constrained submodular optimization-based AL for the scalable histopathological image analysis, which considered the diversity among selected histopathological imaging samples. To the best of our knowledge, only one paper (Hoi *et al* 2006) included both AL and mammographic diagnosis, presenting a framework for 'batch mode active learning' that applied the Fisher information matrix. However, that study merely treated the mammographic image database as one set of data that was used for testing their general methods and did not focus on the particularities of mammographic diagnosis itself.

3. Approach

Three key points in the MDAL process are (1) the modification of the two-view classification and ordinal regression, (2) the redefinition of SQC for two-view binary classification and ordinal regression, and (3) the rank aggregation for combining each involved SQC with self-adaptive weights.

3.1. Problem definition

Regardless of the repeated iteration process, when considering only one iteration t in the MDAL process, the currently unlabeled dataset is denoted as $\boldsymbol{U}^{(t)}$, which stores $|\boldsymbol{U}^{(t)}|$ pairs of two-view mammographic image samples $u_n^{(t)}$ in the form of a feature vector, where $u_{\mathrm{n}}^{(t)} = [x_{CC\,n}^{(t)}, x_{MLO\,n}^{(t)}]$ $(n \in [1,\ldots,|\boldsymbol{U}^{(t)}|]$, and $|.|$ is a function

used to calculate the length of an array). In addition, the existing labeled dataset is denoted $\boldsymbol{A}^{(t-1)}$, obtained from the previous iteration. For each sample in $\boldsymbol{A}^{(t)}$, $\boldsymbol{a}_{\mathrm{m}}^{(t)} = [x_{CC\,m}^{(t)}, x_{MLO\,m}^{(t)}, y_m]$ ($y_m = \{1, -1\}$ for a classification task or $y_m = \{1,2,3,4,5\}$ for an ordinal regression task). With one SQC, the AL algorithm can select c_n pairs of image samples $\boldsymbol{Q}^{(t)}$ with the highest value from $\boldsymbol{U}^{(t)}$ in every iteration and reconstitute a newly labeled dataset $\boldsymbol{A}^{(t)} = \boldsymbol{A}^{(t-1)} \cup \boldsymbol{Q}^{(t)}$ that is used to train a new learning model $h^{(t)}$ and update the SQC for the next iteration $t+1$. The purpose of the AL method is to acquire a high-performance $h^{(t)}$ with the fewest iterations (labeling costs are positively correlated with the number of iterations).

3.2. Modification of the two-view classification and ordinal regression

Because mammographic diagnosis is a two-view learning task, both ordinal regression and classification problems must learn two target functions by searching two optimal hypotheses (h_{CC} and h_{MLO}). For each mammographic image sample ($x = [x_{CC}, x_{MLO}]$), its diagnostic information, including pathological type and histological grade, can be accurately predicted if these two hypotheses provide the same predictions ($h(x) = h_{CC}(\mathrm{x}) = h_{MLO}(\mathrm{x})$). Conversely, if the predictions of one sample x provided by these two hypotheses are different, the predicted result y^* of the hypothesis with the higher probability will be accepted as the decisive result. The revised discriminant function is represented by formula (3):

$$h\left(x\right) = h_v\left(x\right), \text{where } v = \underset{view\in\ CC,\ MLO}{\arg\max}\ \boldsymbol{P}\left(y_{view}^{*}|x, h_{view}\right),\ y_{view}^{*} = h_{view}(x). \tag{3}$$

Then, the problem is translated into a calculation of the predicted probability of each view.

In this paper, a SVM was applied to accomplish the binary classification task, whose discriminant function is represented in formula (4):

$$\mathrm{h}\left(\mathrm{x}\right) = \mathrm{sign}\left(\boldsymbol{w}^{\boldsymbol{T}}\varphi(x) + \boldsymbol{b}\right). \tag{4}$$

Then, the probability $\boldsymbol{P}(y^*|x, h)$ can be replaced by the distance between the sample x and the hyperplane provided from h as the following formula (5):

$$\boldsymbol{P}\left(y^*|x, h\right) \rightarrow \frac{\boldsymbol{w}^{\boldsymbol{T}}\varphi(x) + \boldsymbol{b}}{\|\boldsymbol{w}\|}. \tag{5}$$

The discriminant function of the ordinal regression is represented by expression (6):

$$h(x) = \underset{j=1\ldots r}{\arg\min}\ \left\{ j : \boldsymbol{w}^{\boldsymbol{T}}\varphi(x) + b_j > 0 \right\}, \text{ where } b_r \text{ is set to } +\infty. \tag{6}$$

Its probability $\boldsymbol{P}\ (y^*|x,\ h)$ is equal to the max (π_1, π_2, π_3…...π_r), and each π_j can be calculated using the nonhomogeneous equation (7).

$$\begin{bmatrix} 1 & -\beta(b_1) & -\beta(b_1) & \ldots & -\beta(b_1) \\ 1 & 1 & -\beta(b_2) & \ldots & -\beta(b_2) \\ \vdots & \ldots & \ddots & \ldots & \vdots \\ 1 & \ldots & 1 & -\beta(b_{r-2}) & -\beta(b_{r-2}) \\ 1 & 1 & \ldots & 1 & -\beta(b_{r-1}) \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix} \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{r-2} \\ \pi_{r-1} \\ \pi_r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{7}$$

where $\beta(\cdot)$ is a calculation formula that can be defined as follows: $\beta\ (\mathrm{a}) = e^{a+\boldsymbol{w}^{\boldsymbol{T}}\varphi(\boldsymbol{x})}$.

3.3. Redefinition of SQC for two-view binary classification and ordinal regression

Most SQCs can be denoted as the following formula (8):

$$\boldsymbol{Q}_k^{(t)} = \bigcup_{e=1}^{c-n} \underset{u\in\boldsymbol{U}^{(t)}}{\arg\min^{e}} \left(f_k^{(t)}\left(u\right)\right). \tag{8}$$

The function $x^* = \arg\min_{x\in U}^{e}\ (\mathrm{f}(x))$ in this formula indicates that $\boldsymbol{x}^*$ is equal to the element x in $\boldsymbol{U}$ whose value $f(x)$ is the eth lowest. For our sample selection problem, $f_k^{(t)}\ (.)$ is the kernel function in the SQC used to calculate the score of every unlabeled sample for sample selection according to the existing labeled samples $\boldsymbol{A}^{(t-1)}$. Different AL algorithms have different $f_k^{(t)}\ (.)$ functions. In this paper, the involved SQCs included uncertainty-based and diversity-based criteria, and we defined $f_1^{(t)}\ (.)$ as the kernel function of uncertainty-based SQCs and $f_2^{(t)}\ (.)$ as the kernel function of diversity-based SQCs.

The uncertainty-based SQC seeks to select some unlabeled samples from the unlabeled sample pool $\boldsymbol{U}^{(t)}$ that have the least certainty. The uncertainty-based SQC relies on the learning model h based on the work of previous iterations, and the learning model is accurate only when there is a certain number of labeled samples in $\boldsymbol{A}$.

Accordingly, the uncertainty criterion generally performs well after the middle stage of the AL process (Huang *et al* 2010). In MDAL, which involves two learning models from different perspectives, the kernel function of its corresponding uncertainty criterion may be redefined by formula (9):

$$f_1^{(t)}(u) = \max\left(\boldsymbol{P}\left(y_{CC}^*|u, h_{CC}\right), \boldsymbol{P}\left(y_{MLO}^*|u, h_{MLO}\right)\right), \text{ where } h_{CC},\ h_{MLO} \text{ are trained by } \boldsymbol{A}^{(t-1)} \tag{9}$$

which indicates that one mammographic image can be defined as informative if the confidence of the prediction from its decisive result is low. For the ordinal regression task, h_{CC}, h_{MLO} should be calculated through the formula (6), and h_{CC}, h_{MLO} for the binary classification task can be obtained from the formula (4).

The diversity-based SQC selected diverse samples. The performance of the diversity-based SQC was generally better than that of the uncertainty-based SQC in the early stage of the AL process, but in the late stage, the performance of the uncertainty-based SQC was superior. That is because that performance did not involve the update to the learning model, and its sample selection was only associated with the distribution of all samples (Huang *et al* 2010). The diversity-based SQC in MDAL is defined according to formula (10):

$$f_2^{(t)}(u) = \max_{a \in A^{(t)}} \left(\measuredangle(u^{CC}, a^{CC}) + \measuredangle\left(u^{MLO}, a^{MLO}\right)\right), \text{ where } \measuredangle\left(x_i, x_j\right) = \cos^{-1}\left(\frac{\kappa\left(x_i, x_j\right)}{\sqrt{\kappa\left(x_i, x_i\right)\kappa\left(x_j, x_j\right)}}\right) \tag{10}$$

where κ is the calculation of the Euclidean distance. The formula indicates that an image sample is the most representative if the distance between itself and its nearest point from both the CC and MLO view is the maximum (minimum cosine of the angle).

3.4. The rank aggregation for combining each involved SQC with self-adaptive weight

MDAL is essentially an enhanced version of the MQCAL, thus it is important to select an ICS. Common ICSs in MQCAL, which are used to combine several SQCs, take two forms: a parallel form and a serial form. In parallel-form MQCAL, each involved SQC is combined based on a weighted-sum weighting parameter λ, and the original problem of sample selection is often implemented as an optimization problem, as described below. Serial-form MQCAL employs each SQC to select a certain number $\boldsymbol{c_j}$ of samples from the selection results of previous SQCs in sequence as a multi-layer filter. Expressions of the combination of uncertainty- and diversity-based SQCs are represented by formulas (11) and (12) (Shen *et al* 2004):

$$\boldsymbol{Q}_{parallel}^{(t)} = \bigcup_{e=1}^{c\_n} \arg\min_{u \in \boldsymbol{U}^{(t)}}{}^{e} \left(\lambda f_1^{(t)}(u) + (1-\lambda) f_2^{(t)}(u)\right) \tag{11}$$

$$\boldsymbol{Q}_{serial}^{(t)} = \bigcup_{e=1}^{c\_n} \arg\min_{u \in \boldsymbol{Q}_{uncertainty}^*}{}^{e} \left(f_2^{(t)}(u)\right), \text{where } \boldsymbol{Q}_{uncertainty}^* = \bigcup_{e=1}^{c\_n^*} \arg\min_{u \in \boldsymbol{U}^{(t)}}{}^{e} \left(f_1^{(t)}(u)\right) \tag{12}$$

where the weighting parameter λ and the parameters of the selection number c_n^* in the inner layer may be regarded as the empirical parameters used to balance each involved SQC because each SQC has different mathematics principles and score ranges.

The ICSs used in MDAL were neither the parallel form nor the serial form. Instead, this paper introduces our proposed rank aggregation-based ICS. In each iteration of the AL process, we obtained the rank list $\boldsymbol{R}_k^{(t)}$ and the score list $\boldsymbol{S}_k^{(t)}$ of the currently unlabeled dataset $\boldsymbol{U}^{(t)}$ from each SQC $f_k^{(t)}(.)$, and this rank aggregation-based ICA assumed that the samples that ranked high on the aggregated rank list $\boldsymbol{R}_{agg}^{(t)}$ should be selected for querying labels as formula (13):

$$\boldsymbol{Q}_{RMQCAL}^{(t)} = \bigcup_{th=1}^{c\_n} \arg\min_{u_n^{(t)} \in U^{(t)}}{}^{th} \boldsymbol{R}_{agg}^{(t)}(u_n^{(t)}), \text{ where } \boldsymbol{R}_{agg}^{(t)} = \arg\min_{R^{(t)}} \frac{1}{L}\sum_{k=1}^{L} \omega_k K(\boldsymbol{R}^{(t)}, \boldsymbol{R}_k^{(t)}) \tag{13}$$

where $\omega_k^{(t)}$ is the self-adaptive weight of each $f_k^{(t)}(.)$, and K is the calculation of Kendall's tau or Spearman's footrule distance (Lin 2010).

The steps of our rank aggregation-based ICS in MDAL can be denoted as follows:

Input: the number of sample selections in each AL iteration c_n, the remaining unlabeled dataset $\boldsymbol{U}^{(t)}$

STEP 1: Two score lists of existing unlabeled samples ($\boldsymbol{S}_1^{(t)}$ and $\boldsymbol{S}_2^{(t)}$) can be obtained via $\boldsymbol{U}^{(t)}$ and formula (14):

$$\begin{cases} \boldsymbol{S}_1^{(t)} = [f_1^{(t)}\left(u_1^{(t)}\right), \ldots, f_1^{(t)}\left(u_i^{(t)}\right), \ldots, f_1{}^{(t)}\left(u_{num}^{(t)}\right)] \\ \boldsymbol{S}_2^{(t)} = [f_2^{(t)}\left(u_1^{(t)}\right), \ldots, f_2^{(t)}\left(u_i^{(t)}\right), \ldots, f_2{}^{(t)}\left(u_{num}^{(t)}\right)] \end{cases}, \text{where } u_i^{(t)} \in \boldsymbol{U}^{*(t)}. \tag{14}$$

Then, their corresponding $\boldsymbol{R}_1^{(t)}$ and $\boldsymbol{R}_2^{(t)}$ can also be derived.

STEP 2: $\boldsymbol{S}_1^{(t)}$ and $\boldsymbol{S}_2^{(t)}$ are normalized and sorted from smallest to largest as $\boldsymbol{S}_1^{*(t)}$ and $\boldsymbol{S}_2^{*(t)}$. This paper postulates that a good SQC should maximize the difference between the scores of selected and unselected samples. Then, the different weights ($\omega_1^{(t)}$ and $\omega_2^{(t)}$) behind $f_1^{(t)}(.)$ and $f_2^{(t)}(.)$ in the current iteration may be calculated using the following formula (15):

$$\omega_k^{(t)} = \frac{\min^{c-n+1}(\boldsymbol{S}_k^{*(t)}) - \min^{c-n}(\boldsymbol{S}_k^{*(t)})}{\max(\boldsymbol{S}_k^{*(t)}) - \min^{c-n}(\boldsymbol{S}_k^{*(t)})}, \text{ where } k \in \{1,2\} \tag{15}$$

where function $\min^e(\boldsymbol{S})$ is the value of the element that is the eth lowest in sequence $\boldsymbol{S}$.

STEP 3: The Markov chain can be applied to realize the rank aggregation (Lin 2010). The transition probability matrix $\boldsymbol{TRANMAT}^{(t)}$ must first be calculated, and each element $\boldsymbol{TRANMAT}^{(t)}(i, j)$ is equal to the weighted transition probability of a pair of remaining unlabeled samples $\boldsymbol{P}(u_i^{(t)} \to u_j^{(t)})$, where $i \neq j$, and $i, j \in 1,\ldots, num$), using formula (16):

$$\boldsymbol{TRANMAT}^{(t)}(i,\ j) = \boldsymbol{P}\left(u_i^{(t)} \to u_j^{(t)}\right) = \begin{cases} \frac{1}{num} \cdot I\left(\sum_{k=1}^{2} \omega_k^{(t)} \cdot \left(I\left(\boldsymbol{R}_k^{(t)}\left(u_i^{(t)}\right) > \boldsymbol{R}_k^{(t)}\left(u_j^{(t)}\right)\right)\right) > 0\right) : \ MC1 \\ \frac{1}{num} \cdot I\left(\sum_{k=1}^{2} \omega_k^{(t)} \cdot \left(I\left(\boldsymbol{R}_k^{(t)}\left(u_i^{(t)}\right) > \boldsymbol{R}_k^{(t)}\left(u_j^{(t)}\right)\right)\right) > \frac{1}{2}\right) : \ MC2 \\ \frac{1}{num} \cdot \sum_{k=1}^{2} \omega_k^{(t)} \cdot \left(I\left(\boldsymbol{R}_k^{(t)}\left(u_i^{(t)}\right) > R_k^{(t)}\left(u_j^{(t)}\right)\right)\right) : \ MC3 \end{cases} \tag{16}$$

where MC1, MC2, and MC3 are three different types of kernels of the Markov chain, and $I(.)$ is an indicator function that is equal to one if conditions within the parentheses are satisfied. Then, $\boldsymbol{TRANMAT}^{(t)}(i, i)$ can be obtained from formula (17) after all $\boldsymbol{TRANMAT}^{(t)}(i, j)$ values have been calculated:

$$\boldsymbol{TRANMAT}^{(t)}\left(u_i^{(t)}, u_i^{(t)}\right) = \boldsymbol{P}\left(u_i^{(t)} \to u_i^{(t)}\right) = 1 - \sum_{i \neq j} \boldsymbol{P}\left(u_i^{(t)} \to u_j^{(t)}\right) \tag{17}$$

Step 4: Because our proposed method involved only two SQCs, the above transition probability matrix was often a large, sparse matrix with several 0 elements. To ensure ergodic results for the transition matrix, a tuning parameter t was introduced and treated as follows (18):

$$\boldsymbol{TRANMAT}^{*(t)}(i,\ j) = \boldsymbol{TRANMAT}^{(t)}(i,\ j) \times (1 - tun) + \frac{tun}{num} \tag{18}$$

where tun is typically set to range from 0.01 to 0.15.

Step 5: The stationary distribution of one transition matrix is its principal left eigenvector, which can be computed from a regular power-iteration algorithm after transposing the above matrix. The value of each element in a stationary distribution may be regarded as a Markov chain score of its corresponding samples. Then, we can obtain the final aggregated rank list ($\boldsymbol{R}_{agg}^{(t)}$) of $\boldsymbol{R}_1^{(t)}$ and $\boldsymbol{R}_2^{(t)}$ by ranking the Markov chain scores from large to small. The top c_n samples with high Markov chain scores were collected as $\boldsymbol{Q}^{(i)}$ to query for labels.

We also observed that the computation complexity of the early stage of our proposed AL algorithm was too high (num is large when t is small). Therefore, it was necessary to remove some samples from $\boldsymbol{U}^{(t)}$ before the first step. According to the characteristics of mammographic images, we suggest first selecting samples from the ambiguous mammographic images (whose predictions from the CC and MLO models are different), followed by unambiguous images, as represented in formula (19):

$$\begin{cases} \boldsymbol{U}^{*(t)} = V, \text{ if } \boldsymbol{V} \neq \varnothing \\ \boldsymbol{U}^{*(t)} = \boldsymbol{U}^{(t)}, \text{ else} \end{cases}, \text{ where } \boldsymbol{V} \subset \boldsymbol{U}^{(t)}, \text{ subject to } \forall u \in \boldsymbol{V},\ h_{CC}^{(t-1)}(u) \neq h_{MLO}^{(t-1)}(u) \tag{19}$$

Then, the above formula (14) can be revised to formula (20),

$$\begin{cases} \boldsymbol{S}_1^{(t)} = [f_1^{(t)}\left(u_1^{(t)}\right), \ldots, f_1^{(t)}\left(u_i^{(t)}\right), \ldots, f_1^{(t)}\left(u_{num}^{(t)}\right)] \\ \boldsymbol{S}_2^{(t)} = [f_2^{(t)}\left(u_1^{(t)}\right), \ldots, f_2^{(t)}\left(u_i^{(t)}\right), \ldots, f_2^{(t)}\left(u_{num}^{(t)}\right)] \end{cases}, \text{ where } u_i^{(t)} \in \boldsymbol{U}^{*(t)} \tag{20}$$

which does not contradict the principles of the AL method, but significantly reduces the computation complexity.

Compared with conventional ICSs, this ICS has three advantages. (1) The empirical parameters are no longer necessary. The tradeoff used to balance each involved SQC is adaptable. The SQC with the higher contribution will be assigned a higher weight. (2) Considering that the contribution of one SQC changes with the stage of the entire AL process, the tradeoff is changed from a static to a dynamic state. Then, each SQC may better utilize the respective advantages and realize the advantages' complementation. (3) Our proposal has excellent scalability and generality; any type and number of SQC may be effectively combined; thus, it is ideal for the practical problem here.

3.5. The entire process of MDAL for mammographic diagnosis

The entire process of our proposed MDAL for mammographic diagnosis is illustrated below, a more intuitive display is shown in figure 3.

The entire process of the MDAL algorithm for mammographic diagnosis

Input: The dataset $\boldsymbol{U}^{(0)}$ that contains abundant unlabeled mammographic images with their HOG feature in CC and MLO views and the number of samples selected in each iteration *c_n*.

Repeat

If the number of iterations $t = 0$

Step 1: Randomly select the first batch of unlabeled samples for labeling as $\boldsymbol{Q}^{(0)}$, $\boldsymbol{A}^{(0)} = \boldsymbol{Q}^{(0)}$ and $\boldsymbol{U}^{(1)} = \boldsymbol{U}^{(0)} \backslash \boldsymbol{Q}^{(0)}$

Else

If the purpose of mammographic diagnosis is a binary classification task

Step 1: Train two binary classification models $h_{CC}^{(t-1)}$ and $h_{MLO}^{(t-1)}$ through $\boldsymbol{A}^{(t-1)}$ and formula (4)

Else the purpose of mammographic diagnosis is an ordinal regression task

Step 1: Train two ordinal regression models $h_{CC}^{(t-1)}$ and $h_{MLO}^{(t-1)}$ through $\boldsymbol{A}^{(t-1)}$ and formula (6)

End

Step 2: Select the samples from $\boldsymbol{U}^{(t)}$ as $\boldsymbol{U}^{*(t)}$ through formula (19); then all the samples in $\boldsymbol{U}^{*(t)}$ may be regarded as the most ambiguous.

Step 3: From formulas (20), (9), (10) and (14), we can obtain the rank lists and score lists of $\boldsymbol{U}^{*(t)}$, $\boldsymbol{U}^{*(t)}$, $\boldsymbol{A}^{(t-1)} \rightarrow \boldsymbol{R}_1^{(t)}$, $\boldsymbol{R}_2^{(t)}$, $\boldsymbol{S}_1^{(t)}$ and $\boldsymbol{S}_2^{(t)}$.

Step 4: The weights of diversity and uncertainty are calculated by formula (15). $\boldsymbol{S}_1^{(t)}$ and $\boldsymbol{S}_2^{(t)} \rightarrow \omega_1^{(t)}$ and $\omega_2^{(t)}$

Step 5: Using the weighted Markov chain method (formulas (16)–(18)), we can obtain the weighted aggregated rank list and select the samples with the top *c_n* values as $\boldsymbol{Q}^{(t)}$. *c_n*, $\omega_1^{(t)}$, $\omega_2^{(t)}$, $\boldsymbol{R}_1^{(t)}$, $\boldsymbol{R}_2^{(t)} \rightarrow \boldsymbol{Q}^{(t)}$

Step 6: Request a label of $\boldsymbol{Q}^{(t)}$ from the Oracle; then, $\boldsymbol{A}^{(t)} = \boldsymbol{A}^{(t-1)} \cup \boldsymbol{Q}^{(t)}$ and $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} \backslash \boldsymbol{Q}^{(t)}$.

End

Until: a stopping criterion is applied or $|\boldsymbol{U}^{(t)}| = 0$.

Output: The labeled dataset $\boldsymbol{A}^{(t)}$, which contains far fewer samples than $\boldsymbol{U}^{(0)}$, and the performance of the learning model trained by $\boldsymbol{A}^{(t)}$.

4. Experiment

4.1. Experimental settings

Our experiments were conducted using the digital database for screening mammography (DDSM), a resource of the mammographic image analysis research community containing approximately 2500 cases. The majority of the cases included two images of each breast with their associated information: the assessment of abnormalities (0: incomplete, 1: negative, 2: benign, 3: most likely benign, 4: suspicious, 5: highly suggestive of malignancy) and the projection positions (MLO and CC views), as shown in figure 4 below (Rose *et al* 2006). For the binary classification experiments, 1406 pairs of candidate images were extracted from the DDSM as experimental samples, and each pair contained two small mammographic images of the same tissue but from two views. Of the pairs, 898 were masses, and the remaining 508 were normal tissue. For the ordinal regression experiments, only 1330 pairs of samples were available, with 508 normal, 19 benign, 170 most likely benign, 414 suspicious and 219 malignant. The method for candidate images extraction includes: the image conversion (Chris Rose *et al* 2006), separation of foreground and background based on k-mean and morphological operation, the foreground enhancement through histogram equalization, and the candidate images extraction using therandom forest given in the paper by Kooi *et al* (2016).

As part of the process, 900D HOG features (Dalal and Triggs 2005) were extracted from each image. As a well-designed feature descriptor, a HOG feature can give a good description of local shape information, and has better invariance to changes in transition, rotation, illumination and shadowing, which might be the appropriate choice for mammographic images with variform local structure that are prone to interference. Then, owing to the high dimensions of the HOG feature, it may require further reduction through PCA (Wold *et al* 1987).

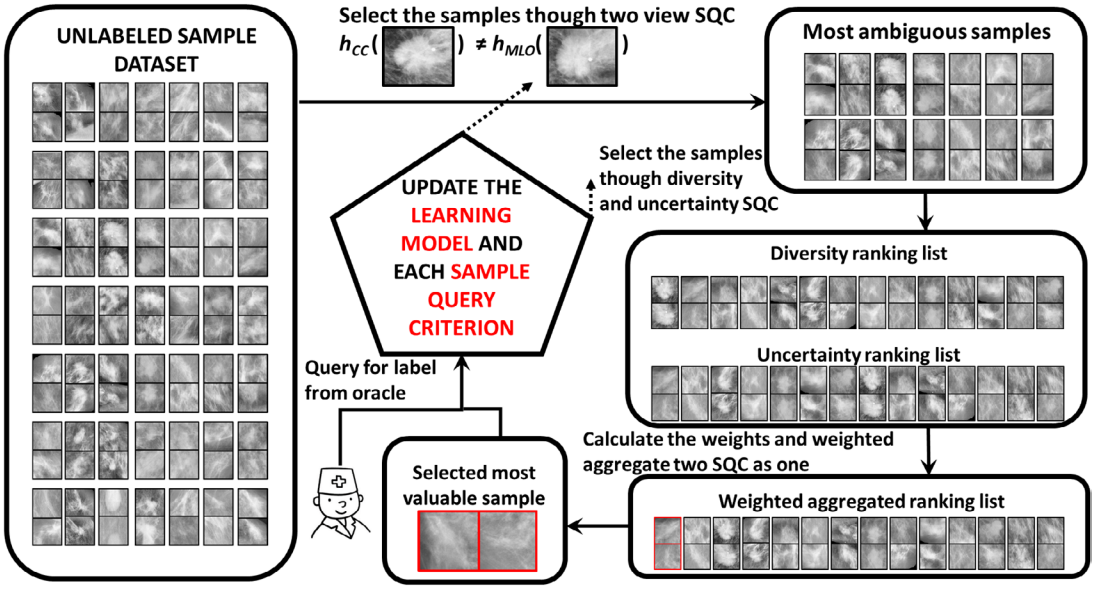

Figure 3. The entire process of our proposed MDAL algorithm for mammographic diagnosis.

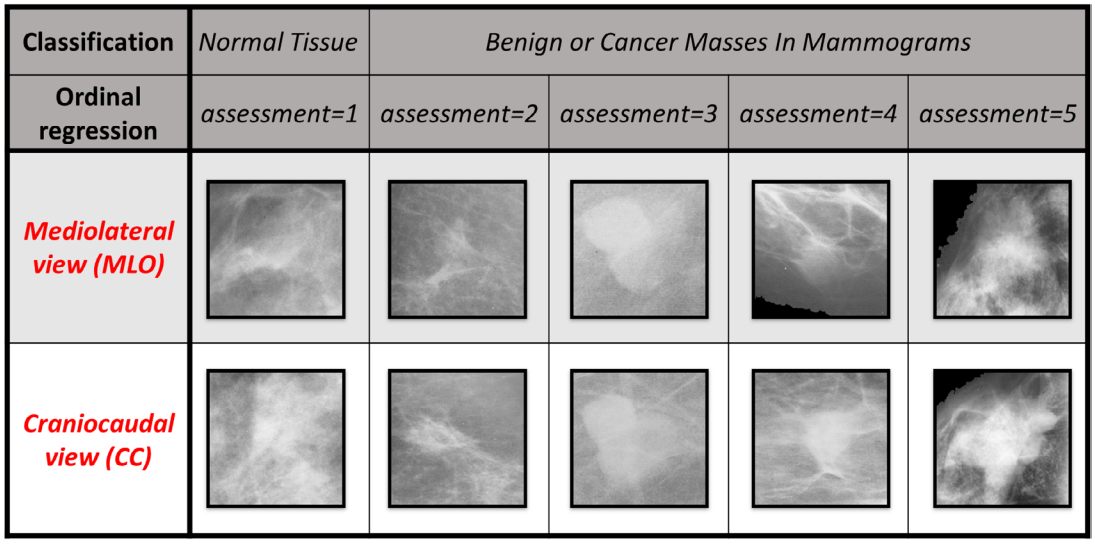

Figure 4. Five pairs of mammographic images with different assessments and labels in DDSM.

It is worth emphasizing that every pair of images is referred to as 'a sample' here, and the labeling costs of a sample is equal to one.

4.2. Experimental environments

All operations were executed using MATLAB R2014a software (The Mathworks, Inc., Natick, MA, USA) installed on a PC with an Intel Core i3-2100 CPU (3.10 GHz) and 3 GB memory. LibSVM supported by www.csie.ntu.edu.tw/~cjlin/libsvm/ was applied to train the SVM classification models with polynomial kernels for all classification experiments. The ordinal regression models were all trained by the built-in function mnrfit.m in MATLAB. All experiments below were repeated a specified number of times. Each time, the corresponding experimental samples were randomly divided into a training set with 50% of the samples and a test set with 50% of the samples, for the classification tasks, and a training set with 80% of the samples and a test set with 20% of the samples for the ordinal regression tasks. It is also worth mentioning that all involved algorithms below were uploaded to GitHub. Any reader interested in this can download them from https://github.com/lestel/MDAL.git.

4.3. Experimental evaluation indexes

For the classification task, AUC (the area under the curve of ROC) and accuracy were used to evaluate the performance of the approaches relative to that described in the paper by Huang *et al* (2010). For the ordinal

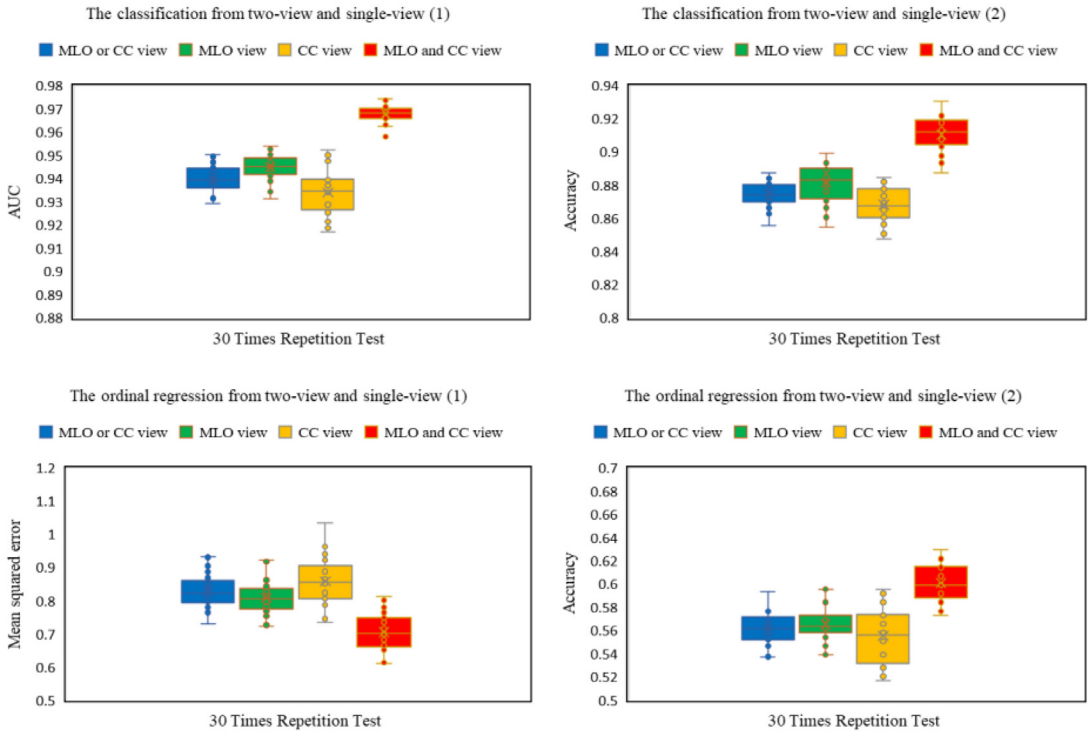

Figure 5. Comparison of the model being trained by single-view or two-view: the figures in the top row are respectively the AUC (left) and accuracy (right) of the classification task, and the figures in the bottom row are respectively the MSE (left) and accuracy (right) of the ordinal regression task.

regression task, this paper uses accuracy and MSE (the mean squared error) as evaluation indexes. A better AL algorithm can help the learning model achieve a higher AUC, accuracy, and a lower MSE with fewer labeling costs.

4.3.1. The learning model for mammographic diagnosis from a single view versus two views

The purpose of this section is to validate the effect on performance if two views are considered in the training process of mammographic classification and the ordinal regression model. In this experiment, we termed the method, which makes the prediction and trains the model from both CC and MLO views, 'CC and MLO views' in the figure below. The other methods for comparison include the 'CC view' or the 'MLO view' (which only considers the mammographic images from one view) and the 'CC or MLO view' (which regards the mammographic images from different views as two unrelated images). All test methods above are used in both a classification and an ordinal regression task. The experiment was repeated 30 times, and the mean and standard deviation of each method were calculated after the repeated tests.

It may be observed from figure 5 that the passive learning classification performance for mammographic diagnosis from both views (AUC $= 0.9680 \pm 0.0038$, accuracy $= 0.9111 \pm 0.0079$) yielded a statistically significant improvement over the classification performance from CC views (AUC $= 0.9339 \pm 0.0093$, accuracy $= 0.8649 \pm 0.0113$), MLO views (AUC $= 0.9448 \pm 0.0057$, accuracy $= 0.8831 \pm 0.096$), and from CC or MLO views (AUC $= 0.9395 \pm 0.0057$, accuracy $= 0.8740 \pm 0.0069$). In addition, the regression performance was also slightly increased for mammographic diagnosis from both views (accuracy $= 0.6011 \pm 0.0158$, MSE $= 0.7054 \pm 0.0571$) compared with the regression performance from CC views (accuracy $= 0.5554 \pm 0.0235$, MSE $= 0.8576 \pm 0.0731$), MLO views (accuracy $= 0.5653 \pm 0.0145$, MSE $= 0.8077 \pm 0.0476$), and from CC or MLO views (accuracy $= 0.5617 \pm 0.0128$, MSE $= 0.8304 \pm 0.0475$). In other words, exploiting the characteristics of two views can reliably improve the performance of the learning model, thus we employed these two-view-based classification and regression models in the following experiments as the baseline of experimental AL algorithms.

4.3.2. AL from two views for classification and ordinal regression in mammography

The experiments in this section were designed to validate that our proposed MDAL can indeed greatly minimize the annotation work in both classification and regression tasks of mammographic diagnosis. The control methods included (1) random, (2) diversity (Demir *et al* 2011), (3) uncertainty (Lewis and Catlett 1994), (4) multiple-view (Muslea *et al* 2006), (5) MCBAL (Shen *et al* 2004), and (6) QUIRE (Huang *et al* 2010). In order to ensure the fairness of the experiments, all these contrasting methods made the following improvements: if one image in any view is selected by them, then this sample is selected. Methods 2, 3, and 4 were the typical

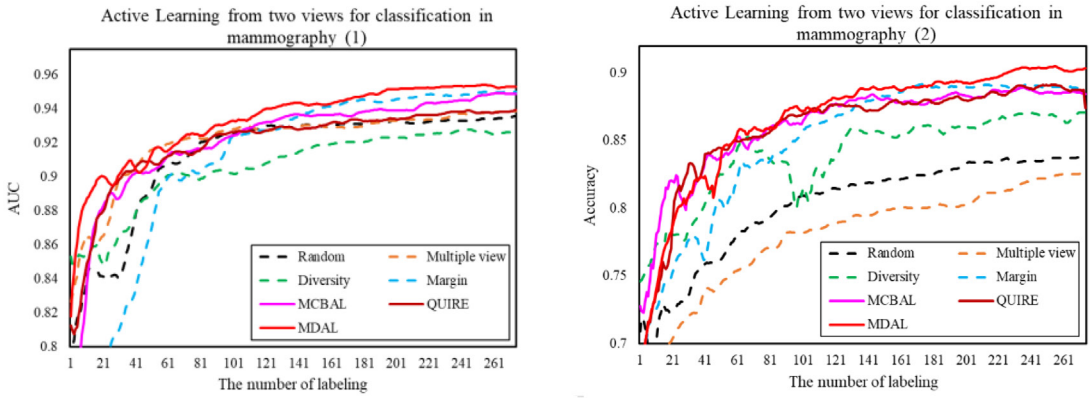

Figure 6. Comparison of seven AL approaches to classification AUC (left) and accuracy (right) with a different number of labels.

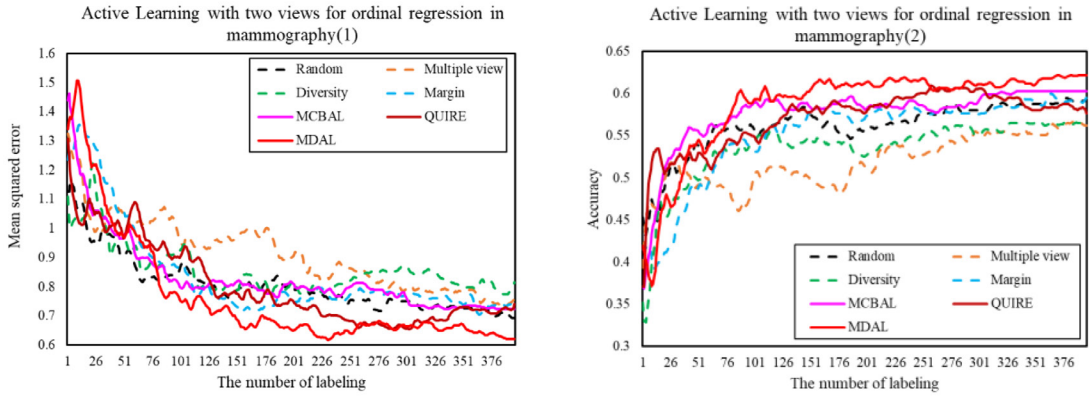

Figure 7. Comparison of seven AL approaches in ordinal regression MSE (left) and accuracy (right) with a different number of labels.

representativeness, informativeness and ambiguity measure-based single criterion AL algorithms, respectively, which were also the components of MDAL. Method 5 was the serial-form MQCAL, and its parameters of the selection number c_n^* was set at 7 here. Method 6 was the parallel-form MQCAL and can be regarded as state of the art, whose weighting parameter λ was equal to 1, as presented in the paper by Huang *et al* (2010). As for MDAL, MC2 was selected as the kernel of the rank aggregation process.

The experiments for each method were repeated 30 times, and the mean and standard deviation of performance were recorded to generate the curves in figures 6 and 7. Before each experiment, we randomly selected four image samples (two masses and two normal tissues) as the initial labeled data set $\boldsymbol{L}^{(1)}$, and only one sample was selected for querying labels in each subsequent iteration ($c\_n = 1$). We also recorded the specific performance values at various important moments for each type of evaluation index (when labeled samples accounted for 5, 10, 15, 20, 30, and 40 percent of all of the training samples at hand), and the best result from each method at these moments was highlighted in bold, based on paired t-tests conducted at the 95 percent level in tables 1 and 3. The comparative results of MDAL, relative to other AL methods for the entire AL process, were also recorded in tables 2 and 4, in which the wins, losses, and ties correspond to the performance of MDAL as above, below, or equal to the other methods in one iteration. In addition, we recorded the average run time of each method to obtain every image that should be labeled in table 5.

The experiment results in figure 6, and tables 1 and 2 elucidate several problems that the majority of the involved methods can address better than random selection, which demonstrates the effectiveness of AL. In particular, our proposed MDAL has performance advantages over diversity, multiple-view, and uncertainty in classification tasks, and these advantages permeate the entire AL process. The reason is that MDAL considers the representativeness, informativeness and ambiguity of each unlabeled sample and makes the best of them. Although the accuracy of QUIRE and MCBAL was slightly better than that of MDAL in the first 40 selected samples, their accuracy advantage was not maintained when more mammographic images were selected.

The results of the ordinal regression task in figure 7, tables 3 and 4 indicate that the performance of single-criterion-based AL, multiple-view, uncertainty and diversity was much worse than that of MDAL and even no better than random selection, which confirmed our view that an appropriate AL algorithm for the

Table 1. The AUC and accuracy comparison of MDAL versus the other AL methods in the classification task of mammography images for labeling fractions of samples (5%, 10%, ...).

	AUC					
	5%	10%	15%	20%	30%	40%
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD
Random	0.85 ± 0.07	0.91 ± 0.01	**0.93 ± 0.01**	0.93 ± 0.01	0.93 ± 0.01	0.94 ± 0.01
Multiple-view	**0.90 ± 0.02**	**0.92 ± 0.01**	**0.93 ± 0.01**	0.93 ± 0.01	0.93 ± 0.01	0.94 ± 0.01
Diversity	0.87 ± 0.02	0.90 ± 0.01	0.90 ± 0.01	0.91 ± 0.01	0.92 ± 0.01	0.93 ± 0.01
Uncertainty	0.83 ± 0.12	**0.90 ± 0.09**	**0.93 ± 0.02**	0.93 ± 0.01	0.94 ± 0.01	**0.95 ± 0.01**
MCBAL	**0.89 ± 0.04**	0.91 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.94 ± 0.01
QUIRE	**0.90 ± 0.02**	0.91 ± 0.01	**0.93 ± 0.02**	0.93 ± 0.02	0.93 ± 0.01	0.94 ± 0.01
MDAL	**0.90 ± 0.01**	**0.92 ± 0.01**	**0.93 ± 0.01**	**0.94 ± 0.01**	**0.95 ± 0.01**	**0.95 ± 0.00**
	Accuracy					
Random	0.75 ± 0.03	0.79 ± 0.03	0.81 ± 0.01	0.82 ± 0.02	0.83 ± 0.02	0.84 ± 0.01
Multiple-view	0.72 ± 0.01	0.76 ± 0.01	0.78 ± 0.02	0.79 ± 0.01	0.80 ± 0.01	0.82 ± 0.01
Diversity	0.79 ± 0.02	**0.85 ± 0.03**	0.82 ± 0.08	0.86 ± 0.09	0.86 ± 0.06	0.87 ± 0.04
Uncertainty	0.78 ± 0.07	0.83 ± 0.04	0.86 ± 0.02	0.88 ± 0.01	**0.89 ± 0.01**	0.89 ± 0.01
MCBAL	0.82 ± 0.09	0.85 ± 0.02	**0.87 ± 0.02**	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.02
QUIRE	**0.83 ± 0.02**	0.85 ± 0.01	**0.87 ± 0.02**	0.87 ± 0.01	0.88 ± 0.01	0.89 ± 0.00
MDAL	0.82 ± 0.01	**0.86 ± 0.02**	**0.88 ± 0.01**	**0.89 ± 0.01**	**0.89 ± 0.01**	**0.90 ± 0.02**

Table 2. Win/tie/loss counts of MDAL versus the other methods in the entire AL process, in the classification task of mammography images.

Classification	AUC			Accuracy		
Task	Wins	Ties	Losses	Wins	Ties	Losses
Random	250	24	0	263	8	3
Multiple-view	184	88	2	272	2	0
Diversity	269	5	0	193	69	12
Uncertainty	147	127	0	183	90	1
MCBAL	217	57	0	153	102	19
QUIRE	207	67	0	164	98	12
In All	1274	368	2	1228	369	47

ordinal regression should have two simultaneous sample query criteria: uncertainty and diversity. Compared with the other two MQCAL, QUIRE and MCBAL, MDAL demonstrated significant advantages in accuracy and MSE, particularly when the number of labels was more than 10% of the total number of unlabeled samples.

From the perspective of running time, because MDAL involves the diversity calculation of some samples, MDAL and diversity were in the same order of magnitude as the running time but far less than QUIRE.

4.4. Clinical tests

We further investigated the number of labels and the actual execution time of the annotation work with or without the assistance of the AL algorithm (here we supposed that similar attention should be paid to each labeled image pair). For this purpose, and in accordance with the principle of the proposed MDAL, a program was established for clinical tests of the AL method for mammographic image annotation, as displayed in figure 8. This program presents the user with one pair of mammographic images from both MLO and CC views in two consecutive manners (MDAL and random), and the user feeds back the labels of these images with no extra hints until the learning model converges (its accuracy exceeds a threshold: 88% for classification, 60% for ordinal regression). In addition, two radiologists from Peking Union Medical College Hospital with two to three years of screening experience in mammographic diagnosis were requested to use this program independently (one for the classification task and the other for the ordinal regression task). Their number of query labeling before the learning model converges (also called labeling cost), average time for sample annotation, the average time of each iteration in the AL process and the total time for mammographic image annotation are recorded in table 6.

Table 3. The MSE and accuracy comparison of MDAL versus the other AL methods in the ordinal regression task of mammography images for labeling fractions of samples (5%, 10%, ...).

	MSE					
	5%	10%	15%	20%	30%	40%
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD
Random	**0.91 ± 0.10**	0.82 ± 0.05	0.78 ± 0.08	0.81 ± 0.06	0.75 ± 0.06	0.69 ± 0.05
Multiple-view	0.92 ± 0.13	0.81 ± 0.08	0.91 ± 0.07	0.81 ± 0.05	0.78 ± 0.05	0.79 ± 0.04
Diversity	0.99 ± 0.12	0.91 ± 0.02	0.78 ± 0.06	0.81 ± 0.05	0.87 ± 0.04	0.82 ± 0.04
Uncertainty	1.07 ± 0.13	0.87 ± 0.12	0.75 ± 0.11	0.75 ± 0.13	0.76 ± 0.05	0.72 ± 0.05
MCBAL	0.98 ± 0.04	0.81 ± 0.07	0.79 ± 0.06	0.82 ± 0.05	0.78 ± 0.05	0.72 ± 0.05
QUIRE	0.96 ± 0.13	0.90 ± 0.07	0.78 ± 0.07	0.76 ± 0.04	**0.67 ± 0.05**	0.74 ± 0.04
MDAL	0.99 ± 0.12	**0.76 ± 0.08**	**0.69 ± 0.05**	**0.65 ± 0.05**	**0.67 ± 0.06**	**0.62 ± 0.06**
	Accuracy					
Random	0.54 ± 0.02	0.57 ± 0.01	0.58 ± 0.02	0.55 ± 0.02	0.58 ± 0.01	0.59 ± 0.01
Multiple-view	**0.57 ± 0.04**	**0.58 ± 0.03**	0.56 ± 0.02	0.58 ± 0.02	0.59 ± 0.01	0.59 ± 0.01
Diversity	0.50 ± 0.05	0.51 ± 0.01	0.55 ± 0.02	0.54 ± 0.02	0.52 ± 0.01	0.54 ± 0.01
Uncertainty	0.48 ± 0.04	0.54 ± 0.04	0.58 ± 0.04	0.57 ± 0.05	0.58 ± 0.02	0.60 ± 0.02
MCBAL	0.55 ± 0.03	**0.59 ± 0.03**	0.59 ± 0.01	0.58 ± 0.01	0.59 ± 0.01	0.60 ± 0.02
QUIRE	0.53 ± 0.03	0.55 ± 0.02	0.58 ± 0.02	0.57 ± 0.01	**0.60 ± 0.02**	0.58 ± 0.02
MDAL	0.54 ± 0.03	**0.59 ± 0.03**	**0.61 ± 0.02**	**0.61 ± 0.01**	**0.61 ± 0.01**	**0.62 ± 0.01**

Table 4. Win/tie/loss counts of MDAL versus the other methods in the entire AL process, in the ordinal regression task of mammography images.

Ordinal	MSE			Accuracy		
Regression Task	Wins	Ties	Losses	Wins	Ties	Losses
Random	309	18	69	325	39	32
Multiple-view	302	44	50	304	54	38
Diversity	312	33	51	349	24	23
Uncertainty	300	79	17	379	12	5
MCBAL	311	37	48	276	67	53
QUIRE	265	95	36	300	62	34
IN ALL	1799	306	271	1933	258	185

Table 5. Comparing the CPU time of MDAL with other methods.

	Random	Multiple-view	Diversity	Margin	MCBAL	QUIRE	MDAL
Classification	0.0125	0.1405	0.8125	0.0313	0.0625	14.731	0.7656
Ordinal Regression	0.0625	0.7813	3.9375	0.1094	0.1094	33.81	3.3844

The results of these clinical tests further demonstrate that the proposed MDAL is beneficial in establishing a learning model for mammographic diagnosis. Comparing with the traditional annotation procedure, where each sample selection for labeling is random, MDAL can shorten the entire time required for sample annotation for both regression and classification tasks. Although the required computation time increased several-fold, the time for the entire annotation procedure with MDAL nevertheless decreased by nearly 53 percent compared to the traditional method in the classification task and nearly 57 percent in the ordinal regression task, because the number of samples required for annotation was drastically reduced.

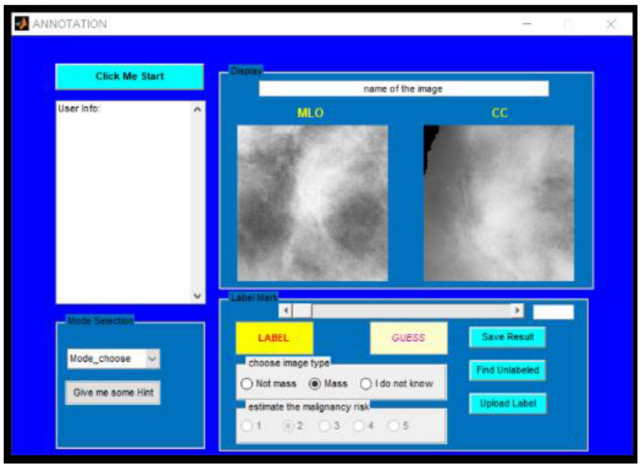

Figure 8. GUI of mammographic image annotation program.

Table 6. Total time spent on mammographic image annotation.

	Labeling cost	Annotation time of doctor (s/each pair)	Time required for each sample (s/each pair)	Total time (s)
Random selection for classification	339	2.2	0.01	339 (2.2 + 0.01) = 749.19
MDAL for classification	118	2.2	0.77	118 (2.2 + 0.77) = 350.46
Random selection for ordinal regression	735	2.9	0.06	735 (2.9 + 0.06) = 2175.6
MDAL for ordinal regression	146	2.9	3.4	146 (2.9 + 3.4) = 919.8

5. Conclusion

In this paper, we proposed a novel AL approach, termed 'MDAL', which was specifically designed for mammographic images.MDAL is different from the AL method designed for general problems as it fully considers the two particularities of mammographic diagnosis. Through its modified discriminant functions, improved sample query criteria and the novel rank aggregation-based MQCAL framework with self-adaptive weightings, MDAL not only further reduced the labeling costs of mammographic images but also extended the AL approach to the ordinal regression task. The experiments on the massive mammographic images from DDSM demonstrated that, compared with other classical AL and state-of-the-art MQCAL methods, MDAL is able to achieve the most optimal results in both the ordinal regression and the classification tasks of mammographic images.

From a clinical perspective, these results are of great significance to the application of MDAL. MDAL can largely eliminate radiologists' burden of annotation work. Numerous unimportant images do not require urgent labeling, nor do those images significantly enhance the learning model. The accuracy of manual annotation can also be improved if we can obtain an effective learning model as soon as possible and utilize its opinion.

Our future work will address the three limitations of this paper. First, we will introduce support vector ordinal regression (a recent improved version of ordinal regression) to our method to further enhance prediction performance although this improvement is independent of the study of AL algorithms. Second, this study is based on the assumption that the radiologist-provided labels are perfect and never incorrect. However, the results of experiments are not perfect; the radiologist also makes mistakes in his annotation work, and thus we will also evaluate the AL method considering imperfect annotations. Third, considering DDSM is an old database, there may exist characteristic differences to the current digital dataset. In the next phase of research, to further verify the validity of the algorithm, we will attempt to modify and apply the proposed MDAL into our local self-developed database that is currently being built.

Acknowledgments

This research is partially supported by the National Key research and development program (2016YFC0106200), the 863-national research fund (2015AA043203), the Chinese NSFC research fund (61190120, 61190124 and 61271318) and the special funding of capital health research and development with No. 2016-1-4011. In addition, the authors of this paper are also grateful to the University of South Florida, which provided the public data sets of mammography in DDSM; Peking Union Medical College Hospital; and the Department of Computer Science at Rutgers University for their personal and technical support.

Conflicts of interest and ethical approval

The authors declare that there is no conflict of interest regarding the publication of this paper. The study does not involve human or animal subjects.

Appendix

Abbreviation

AL	Active learning
SVM	Support vector machine
SQC	Sample query criteria
MLO	The mediolateral view
CC	The craniocaudal view
MDAL	Mammographic diagnosis-based active learning
HOG	Histogram of oriented gradient
TED	Transductive experimental design
MAED	Manifold adaptive experimental design
MQCAL	Multiple query criteria active learning
ICS	Integration criteria strategy
OVR	One-versus-rest
CBA	Classification uncertain area
CCA	Classification compatible area
CNN	Convolutional neural network
AUC	Area under the ROC curve
ROC	Receiver operating characteristic curve
MSE	Mean squared error
MC	Markov chain
DDSM	Digital database for screening mammography

Notations

n	The number of mammographic images
x_i	The ith mammographic image
r	The number of categories
$\boldsymbol{X}$	The feature vector of x, $X = \varphi(x)$
p	Number of feature values in $\boldsymbol{X}$
Xk	The kth feature values in feature vector $\boldsymbol{X}$
π_j	The conditional probability $\mathbf{P}(y = j\|\mathbf{X})$
$\boldsymbol{U}^{(t)}$	The subset of unlabeled samples in tth iteration
$\|.\|$	The length of vector
$u_n^{(t)}$	The $\boldsymbol{n}$th mammographic image in unlabeled subset in tth iteration
t	One iteration of AL process
$x_{CC\,n}^{(t)}$	The $\boldsymbol{n}$th mammographic image in CC view in tth iteration
$x_{MLO\,n}^{(t)}$	The $\boldsymbol{n}$th mammographic image in MLO view in tth iteration
y_m	The label of x_m, $y_m = \{1, -1\}$ for a classification task or $y_m = \{1,2,3,4,5\}$ for an ordinal regression task
$\boldsymbol{A}^{(t)}$	The subset of labeled mammographic images in tth iteration
$a_{\mathrm{m}}^{(t)}$	The mth mammographic image in labeled subset in tth iteration
c_n	Number of images selected for labeling from unlabeled subset in every AL iteration
$\boldsymbol{Q}^{(t)}$	Images selected for labeling from unlabeled subset in tth AL iteration
$h^{(t)}$	The learning model established by $\boldsymbol{A}^{(t)}$ in tth AL iteration
h_{MLO}	The learning model established by images from MLO view
h_{CC}	The learning model established by images from CC view
y^*	The predicted result based on two-view
$f_k^{(t)}(\cdot)$	The kernel function in this SQC that is used to calculate the score of

$x^* = \arg\min_{x \in X^{th}}(f(x))$	x^* is equal to the element x in $\boldsymbol{X}$ whose value $f(x)$ is the th lowest
SQC_k	The kth involved sample query criterion
$\kappa(.,.)$	The kernel distances
λ	Weighted-sum weighting parameter
c_n^*	The selection number in the inner layer of serial-form MQCAL
$\boldsymbol{S}_k^{(t)}$	The scoring list of $u_n^{(t)}$ in tth iteration, given by SQC_k
$\boldsymbol{R}_k^{(t)}$	The ranking list of $u_n^{(t)}$ in tth iteration, given by SQC_k
$\boldsymbol{R}_{agg}^{(t)}$	The aggregated ranking list of $u_n^{(t)}$ in tth iteration, given by all involved SQC_k
$K(.)$	The calculation of Kendall's tau or Spearman's footrule distance
L	The number of involved SQC_k
$\min^i(\boldsymbol{S})$	The ith smallest value in $\boldsymbol{S}$
$\omega_k^{(t)}$	The weight of SQC_k in tth AL iteration
$I(.)$	An indicator function that is equal to one if conditions within the parentheses are satisfied; otherwise, it is equal to zero
$\boldsymbol{S}_k^{*(t)}$	The $\boldsymbol{S}_k^{(t)}$ after being normalized from -1.0 to 1.0 and sorts from smallest to largest
$P(.)$	Probability assignment
$\boldsymbol{Tranmat}\ (i, j)$	Weighted transition probability
tun	One tunable parameters for ensuring ergodic results for the transition matrix, which is usually set to range from 0.01 to 0.15
num	The number of remaining unlabeled samples
$\boldsymbol{w}$	The weights of the prediction model that need to be calculated, where $\boldsymbol{w} = [w1, w2, \ldots, wp]$ for the ordinal regression model and $\boldsymbol{w} = [[w11, w12, \ldots, w1p], [w11, w22, \ldots, w2p], \ldots, [w(r-1)1, w(r-1)2, \ldots, w(r-1)p]]$ for the multiclass classification model.
$\boldsymbol{b}$	The weights of the prediction model that need to be calculated, where $\boldsymbol{b} = [b1, b2, \ldots, b(r-1)]$
$[c1, c2, \ldots, cr]$	r possible categories of samples
$\beta(\mathrm{a})$	The function $\beta\ (\mathrm{a}) = e^{a+\boldsymbol{w}^T\varphi(x)}$

ORCID iDs

Lixu Gu https://orcid.org/0000-0002-6210-4847

References

Cai D and He X 2012 Manifold adaptive experimental design for text categorization *IEEE Trans. Knowl. Data Eng.* **24** 707–19

Chhatwal J, Alagoz O, Lindstrom M J, Kahn C E Jr, Shaffer K A and Burnside E S 2009 A logistic regression model based on the national mammography database format to aid breast cancer diagnosis *Am. J. Roentgenol.* **192** 1117–2

Chris Rose D T, Williams A, Wolstencroft K and Taylor C 2006 Web services for the DDSM and digital mammography research *Int. Conf. on Digital Mammography* pp 376–83

Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2005. CVPR 2005* (IEEE) pp 886–93

Dasgupta S and Hsu D 2008b Hierarchical sampling for active learning *Proc. of the 25th Int. Conf. on Machine Learning* (Helsinki: ACM) pp 208–15

Demir B, Persello C and Bruzzone L 2011 Batch-mode active-learning methods for the interactive classification of remote sensing images *IEEE Trans. Geosci. Remote* **49** 1014–31

Donmez P, Carbonell J G and Bennett P N 2007 Dual-strategy active learning *2007 European Conf. on Machine Learning* (DBLP) pp 116–27

Douak F, Melgani F and Benoudjit N 2013 Kernel ridge regression with active learning for wind speed prediction *Appl. Energy* **103** 328–40

Freund Y, Seung H S, Shamir E and Tishby N 1997 Selective sampling using the query by committee algorithm *Mach. Learn.* **28** 133–68

Gayathri B, Sumathi C and Santhanam T 2013 Breast cancer diagnosis using machine learning algorithms—a survey *Int. J. Distrib. Parallel Syst.* **4** 105

Guo H and Wang W 2015 An active learning-based SVM multi-class classification model *Pattern Recognit.* **48** 1577–97

Gupta S, Chyn P F and Markey M K 2006 Breast cancer CADx based on BI-RAds descriptors from two mammographic views *Med. Phys.* **33** 1810–7

Hoi S C H, Jin R, Zhu J and Lyu M R 2006 Batch mode active learning and its application to medical image classification *Proc. of the 23rd Int. Conf. on Machine Learning* (Pittsburgh, PA: ACM) pp 417–24

Huang S J, Jin R and Zhou Z H 2010 Active learning by querying informative and representative examples *2010 Int. Conf. on Neural Information Processing Systems* pp 892–900

International Agency for Research on Cancer 2012 *GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012* (Lyon: International Agency for Research on Cancer)

Irwig L, Houssami N and van Vliet C 2004 New technologies in screening for breast cancer: a systematic review of their accuracy *Br. J. Cancer* **90** 2118–22

Jiang S and Qing-Yu O U 2015 Batch-mode active learning approach of computer viruses classifier based on information density *J. Naval Univ. Eng.* **4** 31–5

Joshi A J, Porikli F and Papanikolopoulos N 2009 Multi-class active learning for image classification *IEEE Conf. on Computer Vision and Pattern Recognition* pp 2372–9

Kooi T, Litjens G, Van G B, Gubern-Mérida A, Sánchez C I, Mann R, Den H A and Karssemeijer N 2016 Large scale deep learning for computer aided detection of mammographic lesions *Med. Image Anal.* **35** 303

Lewis D D and Catlett J 1994 Heterogeneous uncertainty sampling for supervised learning *Machine Learning Proc. 1994* (New York: Elsevier) pp 148–56

Lin S 2010 Rank aggregation methods *Wiley Interdiscip. Rev. Comput. Stat.* **2** 555–70

Malich A, Fischer D R and Bottcher J 2006 CAD for mammography: the technique, results, current role and further developments *Eur. Radiol.* **16** 1449–60

Muslea I, Minton S and Knoblock C A 2006 Active learning with multiple views *J. Artif. Intell. Res.* **27** 203–33

O'Neill J 2015 An evaluation of selection strategies for active learning with regression *Masters Dissertation* Dublin Institute of Technology

Oliver A, Freixenet J, Marti J, Perez E, Pont J, Denton E R and Zwiggelaar R 2010 A review of automatic mass detection and segmentation in mammographic images *Med. Image Anal.* **14** 87–110

Panda N, Goh K-S and Chang E Y 2006 Active learning in very large databases *Multimedia Tools Appl.* **31** 249–67

Rose C, Turi D, Williams A, Wolstencroft K and Taylor C 2006 Web services for the DDSM and digital mammography research *Int. Workshop on Digital Mammography* pp 376–83

Roy N and McCallum A 2001 Toward optimal active learning through Monte Carlo estimation of error reduction *ICML (Williamstown)* pp 441–8

Schein A I and Ungar L H 2007 Active learning for logistic regression: an evaluation *Mach. Learn.* **68** 235–65

Settles B 2012 Active learning *Synth. Lectures Artif. Intell. Mach. Learn.* **6** 1–114

Settles B, Craven M and Ray S 2008 Multiple-instance active learning *2008 Conf. on Neural Information Processing System: NIPS* pp 1289–96

Shen D, Zhang J, Su J, Zhou G and Tan C-L 2004 Multi-criteria-based active learning for named entity recognition *Proc. of the 42nd Annual Meeting on Association for Computational Linguistics* (Barcelona: Association for Computational Linguistics) p 589

Winship C and Mare R D 1984 Regression models with ordinal variables *Am. Sociol. Rev.* **49** 512–25

Wold S, Esbensen K and Geladi P 1987 Principal component analysis *Chemometr. Intell. Lab. Syst.* **2** 37–52

Yu K, Bi J and Tresp V 2006b Active learning via transductive experimental design *Proc. of the 23rd Int. Conf. on Machine Learning* (Pittsburgh, PA: ACM) pp 1081–8

Zhou Z *et al* 2017 Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally *2017 IEEE Conf. on Computer Vision and Pattern Recognition* pp 7340–9

Zhu Y *et al* 2014 Scalable histopathological image analysis via active learning *2014 Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 369–76