I don't know: Double-strategies based active learning for mammographic mass classification

Jingyang Zhang, Dong Chen, Hongzhi Xie*, Shuyang Zhang, and Lixu Gu*, *Senior Member, IEEE*

***Abstract*—Automatic classification of mammographic mass is an important yet challenging task. Despite the great success of active learning applied to reducing labeling cost, the traditional active learning methods are not suitable for the real-world mammographic mass classification. Because of the uncertain and insufficient knowledge, the radiologist only presents "I don't know" to some queried mammographic cases. To solve this problem, the radiologist's knowledge information is modeled via diverse density concept. Then a novel double-strategies based active learning framework, which is an adaptive combination of the radiologist's knowledge information model and the most uncertainty sampling strategy, together with mutual information based sampling strategy, is developed. Each queried instance is guaranteed not only with high efficiency for training an accurate classifier, but also with high probability of belonging to radiologist's certain knowledge. Experiments on digital database for screening mammography demonstrate that our approach can obtain a reliable classifier for mammographic mass with fewer querying operations compared to traditional active learning methods.**

I. Introduction

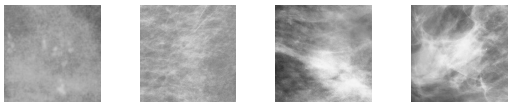
As the major causative factor of breast cancer, mass plays an important role in clinical breast exams. Mammography is one of the most widely used imaging techniques for mass detection. The expensive labeling process, necessary for constructing an automatic mammographic mass classification diagnosis system, can be relieved by active learning [1]. The informativeness measure e.g. the most uncertainty sampling method (MU) [2] serves as a popular selection strategy in traditional active learning framework. But MU only captures the relationship between the queried instance and the labeled instances, ignoring the distributional information contained in a large number of unlabeled instances. This limitation leads to querying outliers and has a negative impact on the generalization performance of the classifier. Therefore, the representativeness measure [3-4] e.g. mutual information based method (MI) is proposed to exploit representativeness behind the unlabeled data and proves beneficial for selecting valuable instances. The combination framework [5] of these two measures shows the outstanding performance of constructing an accurate classifier with fewer but more valuable instances.

*Corresponding author

This research is partially supported by 863 national research fund (2015AA043203) as well as the National Key research and development program (2016YFC0106200).

Jingyang Zhang, Dong Chen and Lixu Gu are with the Laboratory of Image Guided Surgery and Therapy (IGST), Shanghai Jiao Tong University, Shanghai, China (e-mail: J.Y.Zhang@sjtu.edu.cn, chendongyh@126.com gulixu@sjtu.edu.cn,;).

Hongzhi Xie and Shuyang Zhang are with the Department of Cardiothoracic Surgery, Peking Union Medical College Hospital, Beijing, China (e-mail: xiehongzhi@medmail.com, shuyangzhang80@gmail.com)



(a) (b) (c) (d)

Figure 1. Examples of mammographic images labeled uncertainly. (a) and (b) are the images with limited quality. (c) and (d) are the difficult and ambiguous cases with or without mass.

However, these traditional active learning methods are not suitable for the real-world mammographic mass classification task, due to the impractical assumption that oracle has perfect knowledge and can always provide correct labels for each queried instance. As shown in Fig. 1, the radiologist may be uncertain about some cases and only presents "I don't know" for them. Hence, it is necessary for active learning process to handle the issue of imperfect oracle.

Crowdsourcing [6] theory has been applied to this issue which takes advantage of multiple weak oracles subject to different levels of expertise and incorporates their imperfect labels to obtain relatively accurate labels. Based on crowdsourcing theory, some probabilistic models [7-8] have been developed to make quantitative evaluation for these multiple oracles and their corresponding imperfect labels [9] via expectation maximization algorithm. Nevertheless, these crowdsourcing based methods have some drawbacks. They overrate oracle's capability of labeling and oracles are forced to label the uncertain instances. Therefore, some novel active learning frameworks [10-11] have been proposed to take oracle's knowledge domain into consideration. These methods can intelligently avoid querying instances with uncertain knowledge and then, labels of certain instances can be considered correct. However, the single strategy used in this framework suffers from the shortsighted mechanism existing in MU method and hinders the valid construction of classifier.

To address the aforementioned challenge that imperfect radiologist only presents "I don't know" for uncertain instances, we propose a double-strategies based active learning framework for the real-world mammographic mass classification task in this paper. The major contribution includes: (1) to our best knowledge, this is the first work that analyzes the imperfect radiologist in mammographic mass labeling process for active learning. (2) A unified adaptive framework is established to integrate the probabilistic model of oracle's knowledge, which is based on diverse density concept, with two active learning selection strategies. It avoids instances with uncertain knowledge and enhances the validity of queried instances selection. In this way, a reliable classifier can be constructed with fewer querying operations, and time-consuming querying process could be relieved to some extent in clinical labeling.

II. METHOD

Two key components of the proposed active learning method will be introduced in this section: characterization of oracle's knowledge information and double-strategies based active learning framework.

A. Characterization of oracle's knowledge information

Characterization of oracle's knowledge information is a crucial part to handle the issue of imperfect oracle in mammographic mass labeling process. The diverse density concept [11-12] has been utilized for multiple-instance learning and can also be applied to construct a probabilistic knowledge model for oracle.

Given a concept set C representing the knowledge information of oracle, we aim to construct a probabilistic model based on the transformed feature space R_C to estimate $p(x_i \in B^+)$ which denotes the likelihood of instance x_i belonging to oracle's certain knowledge set B^+. Actually, the best approximation for C is the knowledge base set B:

$$\mathrm{C} = \{b_1^+, \cdots, b_p^+, b_1^-, \cdots, b_q^-\} \quad (1)$$

where $B^+ = \{b_1^+, \cdots, b_p^+\}$ represents the instances labeled explicitly and $B^- = \{b_1^-, \cdots, b_q^-\}$ denotes the instances related to the feedback "I don't know". Then the transformed feature vectors $\mathbf{f}_C(B)$ of knowledge base set B in new feature space R_C can be divided into the transformed feature vector of each instance in B:

$$\mathbf{f}_C(B) = [\mathbf{f}_C(b_1^+), \cdots, \mathbf{f}_C(b_p^+), \mathbf{f}_C(b_1^-), \cdots, \mathbf{f}_C(b_q^-)] \quad (2)$$

$\mathbf{f}_C(b_\tau)$ for $b_\tau \in B$ is defined as the conditional probability $\mathrm{p}(c_k|b_\tau)$ for $\forall c_k \in \mathrm{C}$, which is proportional to Gaussian distance between them:

$$\mathbf{f}_C(b_\tau) = [\,\mathrm{p}(c_1|b_\tau), \cdots, \mathrm{p}(c_m|b_\tau)] \quad (3)$$

$$p(c_k|b_\tau) \propto \mathrm{d}(c_k|b_\tau) = e^{-|c_k - b_\tau|^2/\sigma^2} \quad (4)$$

Considering the feature vectors $\mathbf{f}_C(B)$, the corresponding new labels (B) related to oracle's knowledge can be generated:

$$(\mathrm{B}) = [\mathrm{sign}(b_1^+), \cdots, \mathrm{sign}(b_p^+), \mathrm{sign}(b_1^-), \cdots, \mathrm{sign}(b_q^-)] \quad (5)$$

where $\mathrm{sign}(b_\tau)$ indicates whether b_τ belongs to oracle's certain knowledge. Given the transformed feature vectors $\mathbf{f}_C(B)$ and their corresponding labels (B), we can construct a probabilistic classification model $\hbar(\mathbf{f}_C(B), (B))$ to obtain a new candidate instance x_i's probability of falling into certain knowledge B^+:

$$p(x_i \in B^+) = \hbar(\mathbf{f}_C(B), (B))[\mathbf{f}_C(x_i)] \quad (6)$$

Our motivation is to design an active learning selection strategy that prefers instances with high probability of belonging to oracle's certain knowledge based on $p(x_i \in B^+)$. Thus, the negative impact of imperfect oracle can be attenuated.

B. Double-strategies based active learning framework

Besides the probability of belonging to certain knowledge domain, the value of the queried instances is also a critical factor for constructing an effective classifier. So we develop a novel adaptive double-strategies based active learning framework which not only considers the probabilistic model of oracle's knowledge domain, but also utilize MU and MI method to pick the most valuable instances for querying.

The mathematical expectation can be regarded as a valid alternative for the conventional active learning selection strategy in terms of oracle's knowledge information. The objective function of our proposed framework is formulated to select the queried instance with more value:

$$\underset{x_i}{\mathrm{argmax}}\, E(M(x_i)) = \underset{x_i}{\mathrm{E}} \left(H(y_i|x_i; \hbar(L))^\beta d(x_i)^{1-\beta} \right) \quad (7)$$

where

$$H(y_i|x_i; \hbar(L)) = -\sum_{y_i} p(y_i|x_i; \hbar(L)) \log\left(p(y_i|x_i; \hbar(L))\right) \quad (8)$$

$$d(x_i) = I(x_i, X_{U_i}) = H(x_i) - H(x_i|X_{U_i}) \quad (9)$$

$E(M(x_i))$ is the mathematical expectation for selection measure $M(x_i)$. $\hbar(L)$ denotes a classifier trained on labeled set L, from which label y_i of instance x_i can be predicted with the probability $p(y_i|x_i; \hbar(L))$. The weight factor $\beta \in [0,1]$ controls the relationship between the informativeness measure $H(y_i|x_i; \hbar(L))$ and the representativeness measure $d(x_i)$. As mentioned above, $H(y_i|x_i; \hbar(L))$ corresponding to MU method suffers from the shortsighted mechanism, which only analyses a small set of currently labeled instances L and ignores their relationship with abundant remaining unlabeled instances. Thus, the mutual information $I(x_i, X_{U_i})$ acts as the representativeness measure $d(x_i)$ to capture information density between queried instance x_i and the remaining unlabeled instances X_{U_i}. It can be calculated by the difference between x_i 's entropy $H(x_i)$ and conditional entropy $H(x_i|X_{U_i})$ within a Gaussian process framework shown in [4].

The instance x_i belonging to oracle's uncertain knowledge can be regarded beyond oracle's knowledge domain as well as the classifier $\hbar(L)$. So, the conditional entropy with respect to $x_i \in B^-$ can be simplified as:

$$H(y_i|x_i \in B^-; \hbar(L)) = 0 \quad (10)$$

Hence, the objective function (7) can be reformulated as:

$$\underset{x_i}{\mathrm{argmax}}\, E(M(x_i)) = p(x_i \in B^+) H(y_i|x_i; \hbar(L))^\beta d(x_i)^{1-\beta} \quad (11)$$

Intuitively, the maximization of (11) signifies the selection of queried instances with high probability of falling into oracle's certain knowledge and high efficiency for constructing an accurate classifier.

The weight parameter β tunes the trade-off between those two measure terms. Different from the fixed pre-defined parameter in other works, it should be adaptively selected from pre-defined candidate values and dynamically changed during different active learning stages. Motivated by the nonmyopic searching algorithm proposed in [5], the adaptive selection of β from $[0, 0.1, \cdots, 1]$ is equivalent to the selection of queried instances x_i^β from $[x_i^{\beta=0}, x_i^{\beta=0.1}, \cdots, x_i^{\beta=1}]$ obtained via (11) with different β values. Moreover, the selection of queried instances x_i^β can be achieved by minimizing the expected classification error on unlabeled

instances as the approximation for generalization error on unseen instances:

$$x^* = \underset{x \in x_i^{\beta}}{\operatorname{argmin}} \sum_{y} \operatorname{error}(x, y) * p\left(y|x; \hbar(L)\right) \quad (12)$$

where

$$\operatorname{error}(x, y) = \sum_{x_j \in U \setminus x} \left(1 - P\left(\tilde{y}_j | x_j; \hbar(L \cup (x, y))\right)\right) \quad (13)$$

A new classifier $\hbar(L \cup (x, y))$ is retrained by labeled set L and instance-label pair (x, y). $\tilde{y}_j$ is the predicted label for each remaining instance x_j. $\operatorname{error}(x, y)$ denotes the expected classification error obtained from $\hbar(L \cup (x, y))$. Hence, instance x^* is selected and submitted to oracle for updating the target classifier.

C. The flowchart of the proposed method

A flow diagram of the proposed double-strategies based active learning framework is shown in Fig. 2.

During each active learning iteration, uncertain knowledge set and certain knowledge set can be determined related to radiologist's feedback of previous queried instances. Then the characterization of radiologist's knowledge domain is performed to predict the probability of belonging to oracle's certain knowledge for the each candidate instance. Afterwards, the objective function (11) in the proposed framework involves the probability for certain label, informativeness measure and representativeness measure to select the most valuable instance. The selected instance is submitted to the radiologist for its label. It will be added to the certain knowledge set only if the radiologist can provide a certain label, otherwise the instance should be included in the uncertain knowledge set. Finally, the target mammographic mass classifier is updated based on the current certain knowledge set, and then the next active learning iteration is performed in the same way as shown above.

III. EXPERIMENT

We evaluate the effectiveness of the proposed method by comparing it with other conventional active learning methods.

A. Dataset

The experiment is conducted with the digital database for screening mammography (DDSM) [13]. 2941 images centered at normal tissues and 2281 images centered at masses are cropped to build our dataset as shown in Fig. 3. An imperfect radiologist is requested to provide labels representing whether masses exist in the queried mammographic cases. "I don't know" is the only feedback for the uncertain images during the labeling process.

B. Approaches for comparison

We compare the proposed active learning approach with some traditional related methods, including (1) RANDOM, which represents the random sampling strategy; (2) MU [2], the single strategy based MU method without considering oracle's knowledge domain; (3) MI [4], the single strategy based MI method without considering oracle's knowledge domain; (4) DS [5], the double-strategies based method without considering oracle's knowledge domain; (5) MUUK [11], an improved MU method introducing oracle's Uncertain Knowledge information; and (6) MIUK, extending MI method via introducing oracle's Uncertain Knowledge.

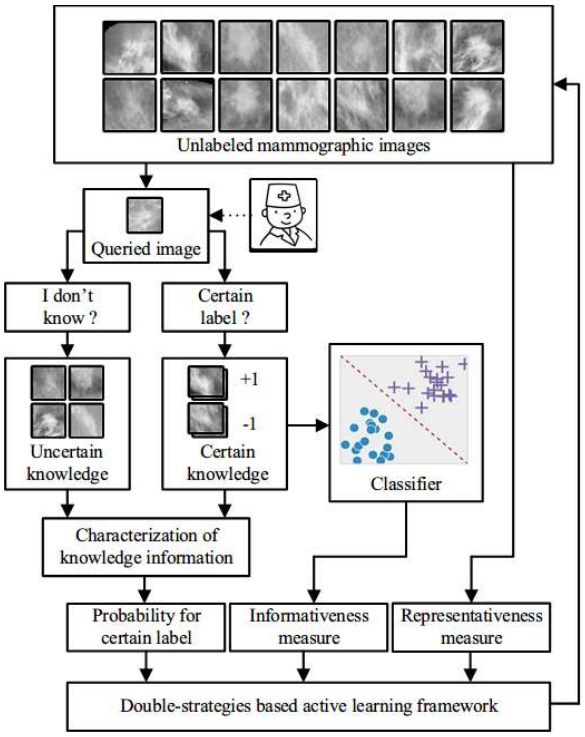
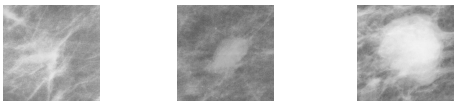


Figure 2. The flowchart of the proposed method

C. Experiment results

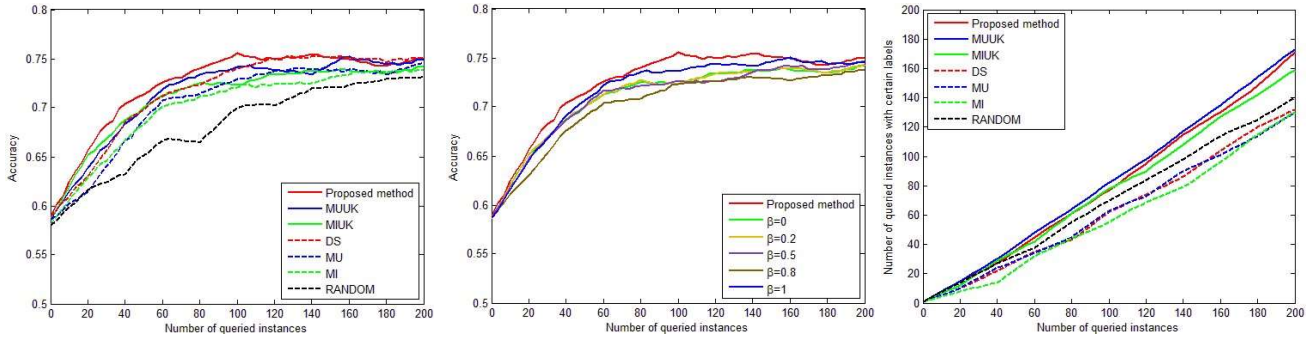
For each analyzed image, original pre-computed 900 HOG features are reduced to 105 features with principal component analysis. Logistic regression classifier acts as the basic classifier to give probabilistic prediction. 75% of the dataset has been randomly picked as training dataset with the remaining regarded as testing dataset. The algorithm randomly selects 50 instances from the training data submitted to the imperfect radiologist as the original labeled set to kick off active learning iteration. The maximal number of iteration is fixed to 200. The active selection process takes about 21 seconds for each case, running the Matlab code on an Intel Core i5-6400 2.70 GHz. The average results of 10-times repeated experiments are shown in Fig. 4.

In Fig. 4(a), the proposed method achieves higher classification accuracy with fewer querying operations compared with all other methods. To reach the convergent classification accuracy of 74.8%, the querying cost of our approach is only 68.4% and 52.5% of that of MUUK and MIUK method. The outperformance over MUUK and MIUK method indicates that the double-strategies framework overcomes the weakness and integrates the strength of these



(a) (b) (c)

Figure 3. Examples in the cropped DDSM dataset. (a) presents the normal tissue. (b) and (c) are related to the benign mass and cancer.



(a) (b) (c)

Figure 4. Numeric results of the experiments. (a) is the comparison result about classification accuracy obtained from different active learning methods. (b) shows the result of the adaptive approach versus the ones using different fixed β. (c) gives the comparison result about the number of certain queried instances.

two single strategies. The similar conclusion can be summarized from the performance of DS over MI and MU method. In addition, Fig. 4(b) verifies the significance of picking adaptive β against using fixed β values. Therefore, the adaptive double-strategies framework contributes to selecting more effective queried instances.

The characterization performance of oracle's knowledge domain is analyzed in Fig. 4(c). The proposed method, MUUK and MIUK can obtain more queried instances with certain labels than the corresponding DS, MU and MI method. This phenomenon proves that the probabilistic model of radiologist's knowledge domain helps the proposed framework avoid querying instances falling into uncertain knowledge. So the construction of classifier can be more valid.

In general, faced with the imperfect radiologist in mammographic mass labeling process, the proposed method demonstrates superior performance compared with other traditional methods.

IV. CONCLSION

In this paper, we develop a novel double-strategies based active learning framework specific to the radiologist's uncertain knowledge during mammographic mass labeling. The diverse density concept based probabilistic model of radiologist's knowledge domain was integrated in adaptive double-strategies active learning framework. Hence, the queried instances can be guaranteed with high efficiency for constructing an accurate classifier, as well as high probability of belonging to radiologist's certain knowledge. Experiments demonstrate that the proposed method can obtain more queried instances with certain labels and achieve a relatively accurate classification result with fewer queried instances compared to other traditional methods. So, the expensive querying operations for constructing an automatic mammographic mass classifier with the real-world imperfect radiologist can be eased to a great extent.

Currently, the target mammographic mass classifier is updated only based on the certain instances. But it is imperfect since some available information behind uncertain instances which may be beneficial for constructing the classifier is ignored. The framework could be extended to refining uncertain labels in the future, with some reliable uncertain labels taken into consideration.

CONFLICT OF INTEREST AND ETHICAL APPROVAL

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

ACKNOWLEDGMENT

This research is partially supported by the National Key research and development program (2016YFC0106200) and 863 national research fund (2015AA043203) as well as the Chinese NSFC research fund (61190120, 61190124 and 61271318).

REFERENCES

[1] B. Settles, "Active Learning Literature Survey," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 2010.

[2] D. D. Lewis, "A sequential algorithm for training text classifiers," *ACM SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

[3] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," *Proc. 25th Int. Conf. Mach. Learn. - ICML '08*, pp. 208–215, 2008.

[4] C. Guestrin, A. Krause, and A. P. Singh, "Near-optimal sensor placements in Gaussian processes," in *Proceedings of the 22nd international conference on Machine learning - ICML '05*, 2005, pp. 265–272.

[5] X. Li and Y. Guo, "Adaptive Active Learning for Image Classification," *Comput. Vis. Pattern Recognit. (CVPR), 2013 IEEE Conf.*, pp. 859–866, 2013.

[6] J. D. Abernethy and R. M. Frongillo, "A Collaborative Mechanism for Crowdsourcing Prediction Problems," *Proc. Adv. Neural Inf. Process. Syst. 24*, pp. 1–9, 2011.

[7] W. Wu, Y. Liu, M. Guo, C. Wang, and X. Liu, "A probabilistic model of active learning with multiple noisy oracles," *Neurocomputing*, vol. 118, pp. 253–262, 2013.

[8] G. Fung, "Active Learning from Crowds," *Proc. 28th Int. Conf. Mach. Learn.*, pp. 1161–1168, 2011.

[9] V. C. Raykar *et al.*, "Learning From Crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, 2010.

[10] D. Tuia and J. Munoz-Mari, "Learning user's confidence for active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 872–880, 2013.

[11] M. Fang and X. Zhu, "Active learning with uncertain labeling knowledge," *Pattern Recognit. Lett.*, vol. 43, no. 1, pp. 98–108, 2014.

[12] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Adv. Neural Inf. Process. Syst.*, pp. 570–576, 1998.

[13] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," *Proc. Fifth Int. Work. Digit. Mammogr.*, pp. 212–218, 2001.