Learning with Unsure Data for Medical Image Diagnosis

Botong Wu[1], Xinwei Sun[2](✉), Lingjing Hu[3,5](✉) and Yizhou Wang[1,4,5]
[1]Computer Science Dept., Peking University
[2] Microsoft Research, Asia
[3] Yanjing Medical College, Capital Medical University
[4] Deepwise AI Lab [5] Peng Cheng Laboratory
{botongwu, yizhou.wang}@pku.edu.cn, Xinwei.Sun@microsoft.com, hulj@ccmu.edu.cn

Abstract

In image-based disease prediction, it can be hard to give certain cases a deterministic "disease/normal" label due to lack of enough information, e.g.*, at its early stage. We call such cases "unsure" data. Labeling such data as unsure suggests follow-up examinations so as to avoid irreversible medical accident/loss in contrast to incautious prediction. This is a common practice in clinical diagnosis, however, mostly neglected by existing methods. Learning with unsure data also interweaves with two other practical issues: (i) data imbalance issue that may incur model-bias towards the majority class, and (ii) conservative/aggressive strategy consideration,* i.e.*, the negative (normal) samples and positive (disease) samples should NOT be treated equally - the former should be detected with a high precision (conservativeness) and the latter should be detected with a high recall (aggression) to avoid missing opportunity for treatment. Mixed with these issues, learning with unsure data becomes particularly challenging.*

In this paper, we raise "learning with unsure data" problem and formulate it as an ordinal regression and propose a unified end-to-end learning framework, which also considers the aforementioned two issues: (i) incorporate cost-sensitive parameters to alleviate the data imbalance problem, and (ii) execute the conservative and aggressive strategies by introducing two parameters in the training procedure. The benefits of learning with unsure data and validity of our models are demonstrated on the prediction of Alzheimer's Disease and lung nodules.

1. Introduction

The early-prediction of disease in medical image analysis often assumes that a deterministic "disease/normal" label can be given to each sample. However, there can be many cases violating such an assumption, since they do not exhibit obvious evidence for determination at early stage. For example, many patients with Alzheimer's Disease (AD) ever experienced an intermediate stage called mild cognitive impairment (MCI) between normal control (NC) and AD. However, MCI does not necessarily convert to AD. The prediction of MCI progression is very difficult at early stage since the MRI/PET of those samples do not show many changes in the lesion regions (e.g. two-side hippo-campus). The accurate prediction for these data needs follow-up examinations, as illustrated by the case in Fig. 2 that it cannot be determined as positive until it converts to AD after 24 months. In contrast, the incautious prediction at early stages for these cases can cause irreversible loss, such as missing golden opportunity for treatment. ***In this paper, we call these cases as "unsure data", which should be considered but has been largely neglected in the literature.***

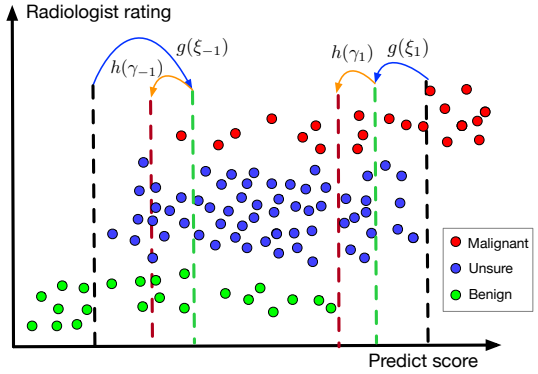

Figure 1. Illustration of our main idea of this work. The red, blue and green points denote the data points of positive, unsure and negative. The points in the region between two dashed lines are predicted as unsure. As shown, the thresholds marked by black are biased towards the majority (unsure) class. Such an effect can be measured by $g(\xi_{\pm 1})$. In addition, there is a gap between the conventional classification and the task in medical image diagnosis, in which the benign class should be predicted with cautious, whereas the malignant should be predicted with high recall. Such a gap can be measured by $h(\gamma_{\pm 1})$. We then introduce cost-sensitive parameters $\xi_{\pm 1}$ and $\gamma_{\pm 1}$ to alleviate imbalanced problem and incorporate conservative/aggressive strategies during learning, respectively.

Table 1. Comparison of predictions between on unsure data and sure data using DenseNet with binary labels on the ADNI dataset.

Data split	# of data	acc	FPR
sure data	24	83.33	13.33
unsure data/MCI	51	61.09	37.04

Most existing methods [16, 14, 4] just simply rule out these unsure data during model training, hence easily result in wrong predictions for this portion of data. To see this, we conducted an experiment on the ADNI dataset[1] of which the MCI class is regarded as unsure class. We train a binary classifier (3D's version of Densenet [13]), regarding AD + MCI-developed-to-AD (MCI_c) as the positive class and the rest (NC + MCI_s) as the negative class. As shown in Table 1, the accuracy and false-positive-rate (FPR) on the unsure data (MCI_c/MCI_s) are much worse than those on the sure data (AD/NC). In other words, the unsure data are hard to be correctly predicted.

There exists one that looks similar, but quite different concept called "hard samples" in the machine learning literature. Hard samples arise either from noisy labels or due to the limitation of model capacity. People try to accurately predict them using active learning [29] or boosting methods [8]. However, unsure data are defined by its nature, *e.g.*, at early stage of disease. Hence, it is hard to give a deterministic label due to lack of information. This type of data can take a large portion of a dataset for model training. We argue that it is responsible and reasonable to identify such unsure data rather than assigning a binary label to each data item without assurance. Labeling a case as unsure practically means that the case needs a follow-up examination.

Compared with the traditional multi-class problem which assumes independence among classes, the "learning with unsure data" faces three challenges: (i) Label dependence issue: the negative (normal), unsure and positive (disease) levels increase in terms of the severity of disease. Hence, the traditional cross-entropy (CE) loss, which assumes the independence among classes, may fail to model such a relationship. (ii) Data imbalance issue: since the unsure data may be the majority class, which may lead the model to bias towards the unsure class, as illustrated by $g(\xi_{\pm 1})$ in Figure 1. (iii) Diagnosis strategy issue: the negative and positive samples are often treated differently in clinical practice, as illustrated by $h(\gamma_{\pm 1})$ in Figure 1. For example, it is reasonable to take recall of malignant cases more seriously than benign ones. Hence, it is important to take such strategy into consideration during diagnosis.

In this paper, we raise the "learning with unsure data" problem and formulate it as an ordinal regression problem and propose a end-to-end learning framework to model the unsure data. In such a framework, three groups of parameters are introduced to address the above three practical issues: threshold parameters, cost sensitive parameters (ξ) and strategy parameters (γ). Specifically, we extend the probability model of binary labels to the classification problem with unsure data by incorporating threshold parameters. To alleviate the data imbalance problem, we further adopt the cost-sensitive loss [17] by introducing cost-sensitive parameters. During the training process, these parameters can be optimized to fit the data from majority class, and hence may lead to more predictions of infrequent classes (positive and negative) with smaller value of threshold parameters. Besides, different from [17], our method can automatically learn the cost-sensitive parameters (together with the threshold parameters and the parameters of backbone neural network) via stochastic gradient descent. Furthermore, to execute the conservative and aggressive (C/A) strategies, we additionally introduce strategy parameters to adjust the margin (threshold) parameters for prediction.

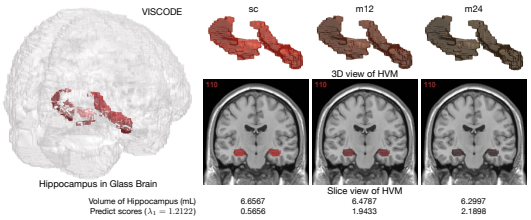

Figure 2. Illustration of a follow-up (screening (sc), 12 month and 24 month) MCI case with Hippocampus in Glass Brain and 3D & slice view of Hippocampus volume mapping (HVM). The volume of HVM is linearly mapped to the color: the darker, the more serious of Hippocampus atrophy. The volume of hippocampus and predicted score by our model are also given.

We apply our model to Alzheimer's Disease (AD) and Lung Nodule Prediction (LNP), in which the early detection is important. For AD, the MCI is regarded as the unsure data and the develop-to-AD/conversion prediction needs follow-up examination. For LNP, we follow the standard in the previous works [4, 26, 15] on Lung Image Database Consortium (LIDC) [22] to label malignant and benign; the others, which are discarded by existing models, are regarded as unsure data. The results demonstrated that our method is superior to others, especially cross-entropy loss. Besides, by considering the data imbalance, the macro-F_1 can be further improved. Moreover, we show that different strategies can result in varying results in terms of precision/recall on positive and negative classes. Particularly, by implementing the C-A strategy, all positive samples are either detected as positive (in most case) or unsure. Such a result agrees with clinical expectations that the negative ones should not miss the opportunity for early treatment. In addition, we also find that learning with unsure data improves the prediction accuracy on sure data.

[1] http://www.loni.ucla.edu/ADNI

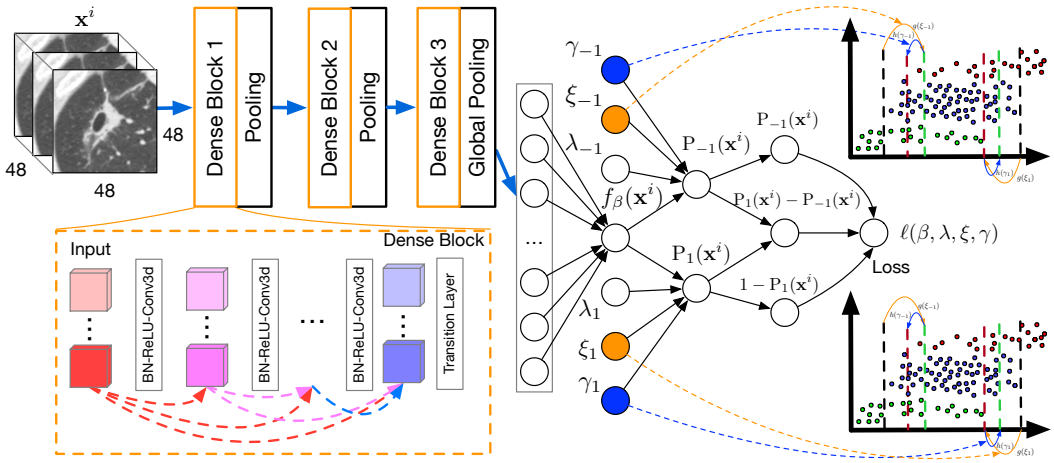

Figure 3. Illustration of the architecture of our unsure data model. A variant of DenseNet is adopted as the backbone. $\xi_{-1,1}$ which are the cost-sensitive parameters denote the orange margin $g(\xi_{-1,1})$ from black threshold to green threshold. $\gamma_{-1,1}$ which are strategy parameters denote the blue margin $h(\gamma_{-1,1})$ from green threshold to red threshold.

2. Related Works

2.1. Modeling Unsure Data

The most related literature considering the unsure data is [28], which considers the partial ranking problem. However, different from partial ranking which considers the pairwise data, the "unsure" data in medical analysis implies that it lacks enough information hence is impossible to determine whether a case/patient is positive (of disease) or not. Moreover, to the best of our knowledge, the unsure data issue has not been explored in medical analysis in the literature. Note that one can not confuse it with hard samples, which have ground truth labels however are easy to be wrongly classified. Correspondingly, many works were proposed to classify those samples, such as active learning [24, 29] and prediction with noisy labels [5].

2.2. Ordinal Regression/Classification

Some works [9, 25] simply cast our task as a common multi-class problem without considering the ordinal relationship among classes, and apply Cross-entropy (CE) loss or mean-variance loss. Besides, [23] transforms ordinal regression as a series of binary classification sub-problems to model the distribution of each sub-problem. However, they ignore the ordinal relationship among negative, unsure and positive classes. Other works regard them as ordinal regression problem, including [19, 10, 20]. In detail, [19] wraps a leading matrix factorization CF method to predict a probability distribution for each class. [10] proposes a probabilistic model with penalized and non-penalized parameters. [20] generates the probabilities for each class via modeling a series of conditional probabilities. [3] uses Poisson and binomial distribution to model each class.

2.3. Imbalance Data Issue

The typical way to alleviate the imbalanced issue [11] is by either over-sampling [6] on minor classes or under-sampling [21] on major classes. Since under-sampling can lose information, [7, 30] presented a way to quantitatively set the weight of minor class based on the cost-sensitive loss. However, the cost matrix should be pre-set. [17] modified the cost matrix in the cross-entropy loss and were able to optimize to learn it iteratively. However, the learning of parameters in cost matrix relies on validation set. In this paper, we modified it to make it useable to ordinal regression loss and can learn it without using validation set.

3. Methodology

Our data consist of N samples $\{\mathbf{x}^i, y^i\}_1^N$ where $\mathbf{x}^i \in \mathcal{X}$ collects the i^{th} sample (e.g. imaging data) and with label $y^i \in \mathcal{Y} = \{-1, 0, 1\}$ with $-1, 0, 1$ denoting the negative, unsure and positive status, respectively. For simplicity, we denote $\mathbf{X}$ and $\mathbf{y}$ as the $\{\mathbf{x}^i\}_1^N$ and $\{y^i\}_1^N$ respectively. The $f_w : \mathcal{X} \to \mathbb{R}$ is a discriminant function (e.g. the neural network output) that is dependent on parameters w.

3.1. Predictive Model with Binary Data

For binary classification problem, the response variable y^i $(i = 1, ..., N)$ are often assumed to be generated from:

$$y^i = \begin{cases} 1 & f_w(\mathbf{x}^i) + \varepsilon^i > 0 \\ -1 & f_w(\mathbf{x}^i) + \varepsilon^i < 0 \end{cases} \tag{1}$$

with $\varepsilon^1, ..., \varepsilon^N \overset{i.i.d}{\sim} \mathrm{G}(\cdot)$. Different G leads to different model and corresponding loss function: (i) Uniform Model: $\mathrm{G}(x) = \frac{x+1}{2}$ (ii) Probit Model: $\mathrm{G}(x) = \Phi(x)$ (Φ is distribution function of $\mathcal{N}(0, 1)$) (iii) Logit Model: $\mathrm{G}(x) = \mathrm{sigmoid}(x) = \frac{\exp(x)}{1+\exp(x)}$. The loss function, which is the

negative log-likelihood of $\mathrm{P}(\mathbf{y}|\mathbf{X}) = \Pi_{i=1}^{N} \mathrm{G}\left(f_w(\mathbf{x}^i)y^i\right)$, is $\ell(w) = -\sum_{i=1}^{N} \log\left(\mathrm{G}\left(f_w(\mathbf{x}^i)y^i\right)\right)$.

3.2. Unsure Data Model

We formulate our problem as ordinal regression problem [19], since the severity of the disease from negative class (-1), unsure data (0) to positive class (1) is increasing. We extend the (1) to incorporate unsure data, by introducing the (1) to model unsure data, with threshold parameter $\lambda \triangleq (\lambda_{-1}, \lambda_1)$ $(\lambda_1 > \lambda_{-1})$,

$$y^i = \begin{cases} 1 & f_w(\mathbf{x}^i) + \varepsilon^i > \lambda_1 \\ 0 & \lambda_{-1} \leq f_w(\mathbf{x}^i) + \varepsilon^i \leq \lambda_1 \\ -1 & f_w(\mathbf{x}^i) + \varepsilon^i < \lambda_{-1} \end{cases} \tag{2}$$

We call such a model as Unsure Data Model (UDM), in which the loss function can be similarly derived as:

$$\begin{aligned} \ell(w, \lambda) = -\sum_{i}^{N} [&1\{y^i = 1\} \log\left(1 - \mathrm{P}_1(\mathbf{x}^i)\right) \\ &+ 1\{y^i = 0\} \log\left(\mathrm{P}_1(\mathbf{x}^i) - \mathrm{P}_{-1}(\mathbf{x}^i)\right) \\ &+ 1\{y^i = -1\} \log\left(\mathrm{P}_{-1}(\mathbf{x}^i)\right)] \end{aligned} \tag{3}$$

where

$$\mathrm{P}_1(\mathbf{x}^i) = \mathrm{G}(\lambda_1 - f_w(\mathbf{x}^i)),\ \mathrm{P}_{-1}(\mathbf{x}^i) = \mathrm{G}(\lambda_{-1} - f_w(\mathbf{x}^i)) \tag{4}$$

During the test stage, the $\mathbf{x}^i$ is simply classified based on the following rule:

$$y^i_{\text{pred}} = \begin{cases} 1 & f_w(\mathbf{x}^i) > \lambda_1 \\ 0 & \lambda_{-1} \leq f_w(\mathbf{x}^i) \leq \lambda_1 \\ -1 & f_w(\mathbf{x}^i) < \lambda_{-1} \end{cases} \tag{5}$$

3.3. Data Imbalance

There are some diseases that are difficult to be employed with accurate prediction, such as lung nodules and Alzheimer's Disease, etc. Hence, it is common for such diseases to have unsure data accounted for a large proportion in the population, which may cause the optimization bias towards the unsure class. To alleviate such a problem, we adopt the idea in [17] by introducing cost-sensitive parameter $\xi \triangleq (\xi_1, \xi_{-1})$ in the training process, in which the $P_{\pm 1}(\mathbf{x}^i)$ in (4) are modified as

$$\begin{aligned} \mathrm{P}_1(\mathbf{x}^i) &= \mathrm{G}\left(-f(\mathbf{x}^i) + \lambda_1 + \log \xi_1\right) \\ \mathrm{P}_{-1}(\mathbf{x}^i) &= \mathrm{G}\left(-f(\mathbf{x}^i) + \lambda_{-1} + \log \xi_{-1}\right), \end{aligned}$$

Note that our method is different from [17]: (1) the parameters in [17] are sample-dependent and ours are only class-dependent; (2) the [17] applies on CE loss and is not applicable to our case; (3) the optimization in [17] relies on validation set, which is not reasonable in medical imaging setting; in contrast, ours can directly implement stochastic gradient descent to optimize ξ, which is simpler and computation efficient.

We explain why such a model can alleviate the data imbalance problem. Note that ξ can partly fit the data to counteract the effect of data imbalance. In more details, the unsure data is often more frequent than sure data in medical analysis. Under such a distribution, directly optimizing $\ell(w, \lambda)$ in (3) may learn large threshold parameter, hence may cause the model to collapse in the unsure data, as shown in the result of cross-entropy in the experimental result. If we set $\log \xi_1 > 0$ and $\log \xi_{-1} < 0$, the learned λ tends to be smaller. During prediction stage, we adopt the same strategy as (5) with threshold parameters $(\lambda_{-1}, \lambda_1)$. With smaller λ, the sure data (infrequent class) will be more encouraged to be successfully classified.

Remark 1 *Note that the UDM, together with data imbalance alleviation, can also be adapted to multi-class classification problem. However, our model mainly focuses on modeling the unsure data, which implies large difficulty to explicitly give a label. Under such cases, it is cautious to classify it as "unsure data", which suggests the follow-up examination or using data from other modalities.*

3.4. Conservative and Aggressive Strategy

In disease prediction of medical analysis, the negative class and positive class are often treated differently in clinical diagnosis. To avoid serious consequences of the disease (such as missing the treatment opportunity), it is expected that the positive samples should be fully detected while control the false-discovery-rate of the negative ones for most diseases, which correspond to aggressive and conservative strategies, respectively. To model such strategies, we propose to introduce parameter $\gamma \triangleq (\gamma_1, \gamma_{-1})$, in which the (4) is modified as

$$\begin{aligned} \mathrm{P}_1(\mathbf{x}^i) &= \mathrm{G}\left(-f(\mathbf{x}^i) + \lambda_1 + \log \xi_1 + \gamma_1\right) \\ \mathrm{P}_{-1}(\mathbf{x}^i) &= \mathrm{G}\left(-f(\mathbf{x}^i) + \lambda_{-1} + \log \xi_{-1} - \gamma_{-1}\right), \end{aligned}$$

Again, since only λ is included during the test stage (i.e., (5)), the enforcement of $\gamma_1 > 0$ may cause smaller value of λ_1, hence can tend to predict positive samples more aggressively. Combining with such a enforcement, the loss function in (3) is then modified as

$$\begin{aligned} g(w, \lambda, \xi, \gamma) = \ell(w, \lambda, \xi, \gamma) + \\ \rho_1 \max(c_1 - \gamma_1, 0) + \rho_{-1} \max(c_{-1} - \gamma_{-1}, 0) \end{aligned} \tag{6}$$

where $c_{\pm 1}$ are preset hyper-parameters. The sign of them correspond to a strategy, i.e.,

Table 2. Comparison between our methods and baselines on LIDC-IDRI dataset. $F_{1,ma}$ (all $\beta = 1$ in Eq. 7) is adopted as evaluation metric.

Loss	$F_{1,ma}$	$Recall_{-1}$	$Recall_0$	$Recall_1$	$Precision_{-1}$	$Precision_0$	$Precision_1$
Poisson[3]	62.59	81.19	37.22	77.34	52.84	68.21	63.87
NSB[20]	66.34	24.31	86.75	74.22	86.89	59.27	68.84
MSE	55.45	93.46	58.25	12.00	92.59	55.05	21.43
CE	65.60	48.17	74.45	69.53	61.05	62.60	78.07
UDM	69.34	20.64	89.91	71.09	88.24	58.64	72.22
UDM+CS	**71.47**	**26.15**	83.91	**79.69**	89.06	59.64	66.67

Table 3. Comparison between our methods and baselines on ADNI dataset. $F_{1,ma}$ (all $\beta = 1$ in Eq. 7) is adopted as evaluation metric.

Loss	$F_{1,ma}$	$Recall_{-1}$	$Recall_0$	$Recall_1$	$Precision_{-1}$	$Precision_0$	$Precision_1$
Poisson[3]	38.17	15.00	95.65	0.00	60.00	58.67	0.00
NSB[20]	36.70	55.00	67.39	0.00	39.29	60.78	0.00
MSE	37.59	10.00	86.96	7.14	33.33	56.34	33.33
CE	32.62	0.00	82.61	50.00	0.00	60.32	41.18
UDM	39.63	50.00	78.26	0.00	47.62	63.16	0.00
UDM+CS	**40.91**	5.00	82.61	**28.57**	33.33	56.72	**40.00**

1. $c_1 > 0, c_{-1} > 0$: conservative strategy on negative class and aggressive strategy on positive class, which matches well with clinical situation.
2. $c_1 < 0, c_{-1} > 0$: conservative strategy on both negative and positive class, which is reasonable in case that the mistakenly diagnose as positive one can also bring serious consequence.
3. $c_1 > 0, c_{-1} < 0$: aggressive strategy on both negative and positive class, which is only applicable to the disease that early detection diagnose is very important.
4. $c_1 < 0, c_{-1} < 0$: aggressive strategy on negative class and conservative strategy on positive class, which is not reasonable in most cases.

We take the first case as an example for explanation. Note that $c_{\pm 1}$ enforces both γ_1 and γ_{-1} to be positive, which encourages smaller value of λ_1 and λ_{-1}. Note that under the same distribution of $\{f(\mathbf{x}^i)\}_{i=1}^N$, combined with the fact that we only use $\lambda_{\pm 1}$ as threshold parameters, the smaller values of both can encourage less and more samples to be classified as negative and positive class, respectively.

Remark 2 *For simplicity, we called conservative on negative and aggressive strategy on positive as c-a strategy, in which the $c_{\pm 1}$ are encouraged to be greater than 0.*

4. Experiments

In this section, we evaluate our models on lung nodule (benign/unsure/malignant) classification and AD/MCI/NC classification. The introduction of datasets (LIDC-IDRI and ADNI dataset) will be introduced in section 4.1, followed by the implementation details and introduction of evaluation metrics in section 4.2 and section 4.3, respectively. We then present our experimental results in section 4.4, and those with conservative-aggressive (c-a) and conservative-conservative (c-c) strategies in section 4.5 and 4.6, respectively. As an explain to the results, we visualize some cases, including bad cases in section 4.7. Finally, we tested our model of predicting sure data in 4.8 to close this section.

4.1. Dataset

For lung nodule classification, we adopt LIDC-IDRI dataset [1], which includes 1010 patients (1018 scans) and 2660 nodules. For each nodule, there are 1-7 radiologists drawing the contour and providing a malignancy rating score (1-5). We followed [4, 14, 27] to label benign and malignant nodules. Specifically, the cases with average score (as) above 3.5 are labeled as malignant; below 2.5 are labeled as benign; others (as: 2.5-3.5) that are dropped by those methods, are labeled as unsure class in our paper.

As for AD/MCI/NC classification task, we adopt ADNI dataset [12], in which samples are MRI images of two-side hippocampus ROIs. As mentioned earlier, the MCI class is regarded as unsure class. The data is split to 1.5T and 3.0T MRI scan magnetic field strength, with 1.5T containing 64 AD, 208 MCI, and 90 NC and 3.0T dataset containing 66 AD, 247 MCI and 110 NC. DARTEL VBM pipeline [2] is then implemented to preprocess the data. The voxel size is $2 \times 2 \times 2$ mm^3 for MRI images.

4.2. Implement Details

All input images of lung nodules and Hippocampus ROIs are cropped as a size of $48 \times 48 \times 48$. We take $G(\cdot)$ as logit model in this paper. Besides, we modified DenseNet [13] as the backbone $(f_w(\cdot))$ for our model. Specifically, we replaced the 2D convolutional layers with the 3D convolutional layers with the kernel size of $3 \times 3 \times 3$. Two 3D convolutional layers with $3 \times 3 \times 3$ kernels are adopted to

Table 4. Comparisons between ours (c-a strategy) and baselines on LIDC-IDRI dataset, with $\beta_1^{\text{rec}} = \beta_{-1}^{\text{pre}} = 2$ in $\text{F}_{\beta,ma}$ in Eq. 7.

Loss	$\text{F}_{\beta,ma}$	Recall_{-1}	Recall_0	Recall_1	Precision_{-1}	Precision_0	Precision_1
Poisson[3]	66.72	66.51	50.47	85.94	61.18	68.97	56.70
NSB[20]	69.77	24.31	86.75	74.22	86.89	59.27	68.84
MSE	60.36	90.09	47.57	44.00	93.59	46.23	26.19
CE	65.55	48.17	74.45	69.53	61.05	62.60	78.07
UDM	69.34	20.64	89.91	71.09	88.24	58.64	72.22
UDM+CS	71.47	26.15	83.91	79.69	89.06	59.64	66.67
UDM+CS+CA	**73.61**	23.85	82.65	**85.94**	**94.55**	60.51	62.86

Table 5. Comparison between ours (c-a strategy) and baselines on ADNI dataset, with $\beta_1^{\text{rec}} = \beta_{-1}^{\text{pre}} = 2$ in $\text{F}_{\beta,ma}$ in Eq. 7.

Loss	$\text{F}_{\beta,ma}$	Recall_{-1}	Recall_0	Recall_1	Precision_{-1}	Precision_0	Precision_1
Poisson[3]	26.88	0.00	47.83	64.29	0.00	53.66	23.68
NSB[20]	30.05	15.00	93.48	0.00	37.50	59.72	0.00
MSE	32.50	10.00	86.96	7.14	33.33	56.34	33.33
CE	32.62	0.00	82.61	50.00	0.00	60.32	41.18
UDM	37.55	0.00	82.61	64.29	0.00	60.32	56.25
UDM+CS	38.38	5.00	82.61	28.57	33.33	56.72	40.00
UDM+CS+CA	**49.41**	25.00	84.78	**35.71**	**45.45**	63.93	62.50

replace the first 2D convolutional layer with the kernel size of 7×7. We used three dense blocks with the size of (6, 12, 24) while the block size of traditional DenseNet121 is (6, 12, 24, 16). To preserve more low-level local information, we discard the first max-pooling layer following after the first convolution layer. We adopt the ADAM [18] with a learning rate of 0.001 to train the network. Restricted by GPU memory, the mini-batch size is set to 4. For data augmentation, we adopted random rotation, shifting and transposing for all training images to prevent overfitting. Both datasets are split into train, validation and test set (1585, 412 and 663 for LIDC-IDRI; 625, 80 and 80 for ADNI). The epoch number is optimized via the performance on the validation set.

4.3. Evaluation Metrics

Since our task belongs to multi-task classification scenario, we adopt metric of $\text{F}_{\beta,ma}$, which is modification of F_β in binary classification and is defined as:

$$\text{F}_{\beta,ma} = \frac{2 \cdot \text{Recall}_{\beta,ma} \times \text{Precision}_{\beta,ma}}{\text{Recall}_{\beta,ma} + \text{Precision}_{\beta,ma}} \tag{7}$$

where

$$\text{Recall}_{\beta,ma} = \frac{\sum_{i=\{-1,0,1\}} \beta_i^{\text{rec}} \text{Recall}_i}{\sum_{i=\{-1,0,1\}} \beta_i^{\text{rec}}}$$

$$\text{Precision}_{\beta,ma} = \frac{\sum_{i=\{-1,0,1\}} \beta_i^{\text{pre}} \text{Precision}_i}{\sum_{i=\{-1,0,1\}} \beta_i^{\text{pre}}}$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \ \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{8}$$

with TP_i, FP_i and FN_i denoting the number of true positive samples, false positive samples, and false negative samples for the i-th class. If all β's in 7 are set to 1, it degenerates to Macro-F_1 ($\text{F}_{1,ma}$), which is often adopted by measuring the performance for multi-class [28]. To evaluate different strategies, we incorporate different β, using weighted precision and recall. In this paper, we mainly consider two strategies by adopting different values of β^{pre} and β^{rec}:

- Conservative-Aggressive (c-a): Focusing more on precision of negative class and recall on positive class, i.e., $\beta_{-1}^{\text{pre}} = \beta_1^{\text{rec}} > 1$, others $= 1$
- Conservative-Conservative (c-c): Focusing more on precision of both negative and positive class, i.e., $\beta_{-1}^{\text{pre}} = \beta_1^{\text{pre}} > 1$, others $= 1$.

4.4. Comparisons with Baselines

To validate the effectiveness of our model, we compared our model UDM with densenet using CE loss, NSB [20], Possion model in [3]. Besides, we also compare another method, which regards it as regression problem. Although it considers the order among three classes, it assumes the increment between consecutive classes are the same, which may not agree with reality. As shown in Table. 2 and Table. 3, our model outperforms others a lot in both tasks. Moreover, by additionally adopting cost-sensitive (CS) parameters (UDM+CS), the result can be further improved and the imbalanced issue can be alleviated. To see this, note that on LIDC-IDRI dataset (Table 2), compared with UDM without CS, the recall of both negative and positive improved 5.51 % and 8.6 %, respectively. Particularly, on ADNI dataset (Table 3), most methods seriously suffer the imbalanced issue. Without CS, the recall of positive for Poisson, NSB, MSE and UDM is almost 0, so as to the negative class of

Table 6. Comparisons between ours (c-c strategy) and baselines on LIDC-IDRI dataset, with $\beta_1^{\text{pre}} = \beta_{-1}^{\text{pre}} = 2$ in $\text{F}_{\beta,ma}$ in Eq. 7.

Loss	$\text{F}_{\beta,ma}$	Recall_{-1}	Recall_0	Recall_1	Precision_{-1}	Precision_0	Precision_1
Poisson[3]	60.35	78.44	35.02	73.44	50.29	62.36	64.83
NSB[20]	67.39	24.31	86.75	74.22	86.89	59.27	68.84
MSE	55.58	93.46	58.25	12.00	92.59	55.05	21.43
CE	67.25	21.10	84.54	78.91	90.20	58.39	66.01
UDM	66.58	35.32	84.86	64.84	79.38	59.78	71.55
UDM+CS	67.87	33.94	81.07	79.69	79.57	61.34	67.55
UDM+CS+CC	**68.72**	29.82	87.38	72.66	**86.67**	60.61	**70.99**

Table 7. Comparison between ours (c-c strategy) and baselines on ADNI dataset, with $\beta_1^{\text{pre}} = \beta_{-1}^{\text{pre}} = 2$ in $\text{F}_{\beta,ma}$ in Eq. 7.

Loss	$\text{F}_{\beta,ma}$	Recall_{-1}	Recall_0	Recall_1	Precision_{-1}	Precision_0	Precision_1
Poisson[3]	36.30	15.00	95.65	0.00	60.00	58.67	0.00
NSB[20]	30.88	15.00	93.48	0.00	37.50	59.72	0.00
MSE	36.25	10.00	86.96	7.14	33.33	56.34	33.33
CE	28.96	10.00	93.48	0.00	33.33	58.11	0.00
UDM	36.39	50.00	78.26	0.00	47.62	63.16	0.00
UDM+CS	39.68	5.00	82.61	28.57	33.33	56.72	40.00
UDM+CS+CC	**42.03**	25.00	76.09	21.43	**35.71**	59.32	**42.86**

Table 8. Comparison between UDM+CS+CA and the binary classifier without unsure data in terms of prediction on sure data in AD and Lung nodule (LN). "R","P" stand for Recall and Precision.

Task	Method	$\text{F}_{1,ma}$	R_{-1}	R_1	P_{-1}	P_1
AD	Binary	79.13	95.00	57.14	76.00	88.89
	Ours	**84.95**	95.00	**71.43**	**82.61**	90.91
LN	Binary	88.54	88.53	89.84	93.69	82.14
	Ours	**88.92**	84.86	**95.31**	**96.85**	78.71

CE. By leveraging CE parameters, our UDM+CS performs much more balanced, as highlighted by the blue color.

To further validate the contribution of CS parameters to alleviating the imbalanced problem, we compare the threshold parameters and also the prediction number of each class. As shown in Fig. 5, the threshold parameters learned by UDM+CS (green lines) are with smaller magnitude than those in UDM (black lines), hence can avoid the model to bias towards the unsure one (the number of cases that are predicted as unsure class: 446 (with CS) v.s. 467).

4.5. Conservative-Aggressive Strategy

We compare c-a strategy with others in this section. For our model 6, $c_1 = c_{-1} = 0.01$ (UDM+CS+CA). To better measure the c-a strategy, we set $\beta_{-1}^{\text{pre}} = \beta_1^{\text{rec}} = 2$ in $\text{F}_{\beta,ma}$ (Eq. 7) for c-a strategy. The $\rho_{\pm 1} = 20$ for LIDC-IDRC and $= 30$ on ADNI dataset, respectively. As shown in Table. 4 and 5, UDM+CS+CA outperforms UDM+CS by 2.14 % on LIDC-IDRI dataset and 11.03 % on ADNI dataset in terms of $\text{F}_{\beta,ma}$. In addition, in terms of Recall_1 and Precision_{-1}, UDM+CS+CA boosts 6.25 % and 5.49 % on for LIDC-IDRI dataset; 7.14 % and 12.12 % on ADNI dataset, compared with UDM+CS. Although Poisson [3] and CE have higher Recall_1 than ours, they bias largely towards positive and unsure classes ($\text{Precision}_{-1} = \text{Recall}_{-1} = 0\%$). Besides, **the positive class is predicted as either positive or unsure, which suggests treatment or more examinations, hence can avoid the irreversible loss.**

Such a improvement can be contributed to the parameters $\gamma_{\pm 1}$. Again, as shown by UDM+CS+CA from Fig. 5, the smaller threshold parameter (λ_1) for positive class encourages more cases to be predicted as positive class (185 v.s. 153), to the aggressive strategy. On the other hand, the smaller threshold parameter (λ_{-1}) for negative class encourages less cases to be predicted as negative class ones (59 v.s. 64), which corresponds to the conservative strategy.

4.6. Conservative-Conservative Strategy

For c-c strategy, we set $c_1 = -0.01$ and $c_{-1} = 0.01$. The $\beta_{-1}^{\text{pre}} = \beta_1^{\text{pre}} = 2$ in $\text{F}_{\beta,ma}$ (Eq. 7). As shown in Table. 6 and 7, UDM+CS+CC outperforms UDM+CS by 0.85 % on LIDC-IDRI dataset and 2.35 % on ADNI dataset in terms of $\text{F}_{\beta,ma}$. Again, UDM+CS+CA improves by 3.44 % on Precision_1 and 7.10 % on Precision_{-1} for LIDC-IDRI dataset; 2.86 % on Precision_1 and 2.38 % on Precision_{-1} on ADNI dataset. Such a result validates the effectiveness of our models to control policy into our model.

4.7. Visualization

We visualize some lung nodules in Fig. 4.7 as an example to illustrate the advantages over others. Top left, top right, bottom left, bottom right correspond to benign (negative), malignant (positive), unsure and some bad cases. The nodules are marked by green box, and the scores below the figure are predicted probability of listed models. Compared with benign nodules, the malignant nodules are with larger size, lower density (more dark), more irregular (e.g. lobu-

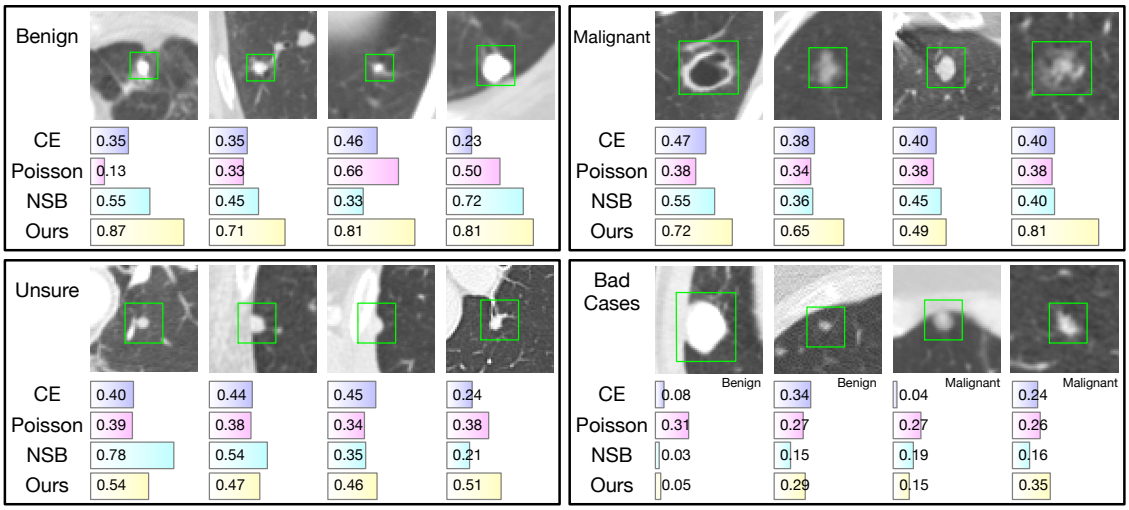

Figure 4. Comparisons between our method (UDM+CS+CA) and baselines for lung nodule prediction in terms of probabilities (scores below each case) of ground truth class. Top-left, top-right, bottom-left, bottom-right subfigures refer to benign, malignant, unsure nodules, and bad cases, respectively. The nodules are marked by green boxes.

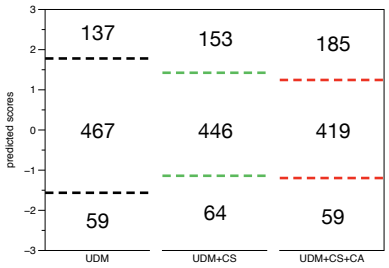

Figure 5. Comparisons in terms of threshold parameters and predicted number of each class among three methods: UDM (black), UDM+CS (green), UDM+CS+CA (red).

lation and spiculation), and connect with vessels or pleura. As shown, the listed benign nodules are with high density and smooth boundary, which can be successfully predicted with high probability while others fail. For malignant ones, they are irregular and with low density. For example, the last one is a large part-solid nodule (PSN) and hence has high malignant degree. Our model predict it as benign with high probability (0.81) while other methods predict as about 0.40. The unsure nodules share part of those characteristics: e.g. (i) the second and third one are close to the pleura, and (ii) the fourth one is irregular. It is unclear to determine whether they belong to benign or malignant, and they are predicted by our model as unsure.

We also listed some hard cases. Again, they are different from unsure class which does not have explicit labels. As shown in the bottom right, there are with explicit labels but hard to classify. For instance, (i) the first benign nodule possesses larger size whereas smooth boundary and high density, which is a typical benign nodule; however, they are close to the pleura. (ii) The first malignant nodule is with pleura indentation, which is a malignant attribute, but it possesses small size. However, all methods, include ours, fail to predict them. We leave the resolvement of the limitations on hard samples in future work.

4.8. Prediction on Sure Data

In this section, we test the ability of our model on predicting the sure data (in both validation and test sets). As an comparison, we also implemented a binary classification model without unsure data in the training process. We can observe from Table. 8 that UDM+CS+CA can outperform the binary model in terms of $\mathrm{F}_{1,ma}$ (with only negative and positive classes). Particularly, the recall on positive class (R_1) and precision on negative class (P_{-1}) largely improve, which can be contributed to the c-a strategies.

5. Conclusion

In this paper, we introduced "unsure data" in medical imaging analysis. We proposed a new framework to model such data and alleviate the effect of imbalanced data. Moreover, we leveraged the conservative and aggressive strategies into our framework in the training procedure. Experiments on lung nodule prediction and AD/MCI/NC classification show that our method outperforms others in terms of performance and interpretability.

6. Acknowledgements

This work was supported in part by NSFC grants 61625201 and 61527804, and Qualcomm University Research Grant.

References

[1] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Hen-

schke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 4.1

[2] J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007. 4.1

[3] C. Beckham and C. Pal. Unimodal probability distributions for deep ordinal classification. In *International Conference on Machine Learning*, pages 411–419, 2017. 2.2, 2, 3, 4, 5, 4.4, 6, 7, 4.5

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 2018. 1, 1, 4.1

[5] R. M. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008. 2.1

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2.3

[7] C. Elkan. The foundations of cost-sensitive learning. 17(1):973–978, 2001. 2.3

[8] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996. 1

[9] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008. 2.2

[10] A. E. Gentry, C. Jacksoncook, D. E. Lyon, and K. J. Archer. Penalized ordinal regression methods for predicting stage of cancer in high-dimensional covariate spaces. *Cancer Informatics*, 2015:201–208, 2015. 2.2

[11] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008. 2.3

[12] C. Huang, X. Sun, J. Xiong, and Y. Yao. Split lbi: An iterative regularization path with structural sparsity. In *Advances In Neural Information Processing Systems*, pages 3369–3377, 2016. 4.1

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 4.2

[14] S. Hussein, K. Cao, Q. Song, and U. Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International conference on information processing in medical imaging*, pages 249–260. Springer, 2017. 1, 4.1

[15] S. Hussein, K. Cao, Q. Song, and U. Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International Conference on Information Processing in Medical Imaging*, pages 249–260. Springer, 2017. 1

[16] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci. Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process. *arXiv preprint arXiv:1703.00645*, 2017. 1

[17] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks*, 29:1–15, 2018. 1, 2.3, 3.3

[18] D. P. Kingma and J. L. Ba. Adam: Amethod for stochastic optimization. In *Proc. 3rd Int. Conf. Learn. Representations*, 2014. 4.2

[19] Y. Koren and J. Sill. Ordrec: an ordinal model for predicting personalized item rating distributions. pages 117–124, 2011. 2.2, 3.2

[20] X. Liu, Y. Zou, Y. Song, C. Yang, J. You, and B. V. Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *European Conference on Computer Vision*, pages 335–344. Springer, 2018. 2.2, 2, 3, 4, 5, 4.4, 6, 7

[21] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009. 2.3

[22] M. F. Mcnittgray, S. G. A. Iii, C. R. Meyer, A. P. Reeves, G. Mclennan, R. C. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, and P. H. Bland. The lung image database consortium (lidc) data collection process for nodule detection and annotation. *Academic Radiology*, 14(12):1464–1474, 2007. 1

[23] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 2.2

[24] R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011. 2.1

[25] H. Pan, H. Han, S. Shan, and X. Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018. 2.2

[26] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017. 1

[27] B. Wu, Z. Zhou, J. Wang, and Y. Wang. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In *ISBI*, pages 1109–1113. IEEE, 2018. 4.1

[28] Q. Xu, J. Xiong, X. Sun, Z. Yang, X. Cao, Q. Huang, and Y. Yao. A margin-based mle for crowdsourced partial ranking. *acm multimedia*, pages 591–599, 2018. 2.1, 4.3

[29] S. Yan, K. Chaudhuri, and T. Javidi. Active learning from noisy and abstention feedback. pages 1352–1357, 2015. 1, 2.1

[30] Z. Zhou and X. Liu. On multi-class cost-sensitive learning. *computational intelligence*, 26(3):232–257, 2010. 2.3