# Optimizing Specificity under Perfect Sensitivity for Medical Data Classification

Cho-Yi Hsiao, Hung-Yi Lo, Tu-Chun Yin and Shou-De Lin
Department of Computer Science and Information Engineering
National Taiwan University
Taipei 106, Taiwan

***Abstract*—One of the main purposes of a computer-aided diagnosis (CAD) system is to reduce the workload of the radiologists in identifying potential diseases. However, such system can become unreliable and useless if it produces even only a small amount of false negatives, since a misclassification of any unhealthy patient as healthy can result in the delay of treatment, which can lead to fatal outcomes. Designing a CAD system that is capable of reducing the workload of radiologists and meanwhile avoiding any false negative is a very challenging problem. To tackle this problem, we propose a two-stage framework and a novel evaluation criterion, namely *optimal specificity under perfect sensitivity* (OSPS). We argue that for medical data classification, this criterion is more suitable than other conventional measures such as accuracy, f-score, or area-under-ROC curve. We further propose two learning strategies to improve OSPS. The first aims particularly at multi-instance learning tasks via disregarding the misclassified negative instances of positive patients. The second tries to improve OSPS by embedding more restricted constraints for negatives.**

## I. Introduction

Computer-aided diagnosis (CAD) systems are designed to assist radiologists in interpreting medical data. CAD systems have widely used for many diseases, such as Alzheimer's disease detection from single photon emitting computer tomography [1] and nodule detection from lung computed tomography [2]. The design of a CAD system can be further decomposed into three stages: Identifying potentially unhealthy regions of interest (ROI), extracting descriptive features for each ROI, and designing a classifier to identify the labels of newly added candidates. This study focuses on the final stage of a CAD system.

Without the help of CAD systems, radiologists generally have to rely on their own eyes and go through all the images or videos to identify potentially unhealthy ROIs. To ensure that no plausible indicators for diseases escape detection, in many cases the second reading on negative data (i.e. the ones believed to be healthy by the first radiologists) by another radiologist is required at the cost of doubling of the workload.

The major goal of a CAD system is to reduce the workload of the radiologists. Some studies have suggested the use of CAD systems as a filter for positive instances in clinical trials [3]. To serve such a purpose, we argue that CAD systems should aim at classifying a *complete negative set*, within which all instances are highly likely to be true negative (i.e. a negative data point that is correctly identified as negative). In other words, a complete negative set contains no false negatives (i.e. a positive data point that is wrongly identified as negative). Guaranteeing a complete negative set can reduce the workload of the radiologists since they do not need to check this set anymore.

Compared to other kinds of data, medical data possesses several special characteristics which in many cases amplify the intricacy of the mining task [4]:

- *Imbalanced data.* For many diseases (in particular for cancer), the positive collections are far fewer than the negative ones.
- *Multiple-instance data.* Medical data are usually bundled together as sets. Moreover, in many cases we only care about the label for that set. A set is positive if at least one of its instances is positive; otherwise it is believed to be negative. Diverse Density [5] and EM-DD [6] are two of the most popular approaches for multi-instance learning.
- *Extremely high penalty for false negatives.* For medical data, failing to identify positive individuals can result in fatal costs while false alarms are generally not as serious. It is different from some other problems (e.g. search engine) where false positives are as serious (if not more severe) than false negatives. The emphasis on false negatives also applies to homeland security, crime analysis and fraud detection data.

This study tries to address the third issue given the first two conditions for medical data mining. To our knowledge, this is the first report that aims at solving the problem of identifying a complete negative set for medical data.

The major contributions are threefold:

1) We propose a framework that divides the task into two sub-tasks: instance ranking and rank-list thresholding. In this study, we focus on tackling the former task.
2) We propose a novel performance measure, the *optimal specificity under perfect sensitivity (OSPS)*, for evaluating how well a *ranking model* benefits a CAD-like system.
3) To improve OSPS for the instance ranking task, we propose two strategies. The *false positive tolerance strategy* and the *suspicion expansion strategy*. The experiments show that our strategies can outperform the state-of-the-art multi-instance models by as much as 20% in OSPS.

## II. Problem Analysis

### A. Preliminaries

A binary classified instance can generally be assigned to one of these following categories:

1) True positive (TP): A positive data point that is correctly identified as positive.
2) False positive (FP): A negative data point that is wrongly identified as positive.
3) False negative (FN): A positive data point that is wrongly identified as negative.
4) True negative (TN): A negative data point that is correctly identified as negative.

Some evaluation metrics have been proposed:

1) Precision $= \frac{\text{TP}}{\text{TP+FP}}$
2) Negative predictive value $= \frac{\text{TN}}{\text{TN+FN}}$
3) Sensitivity (Recall) $= \frac{\text{TP}}{\text{TP+FN}}$
4) Specificity $= \frac{\text{TN}}{\text{TN+FP}}$
5) Accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$

In addition to these evaluation metrics, the receiver operating characteristic (ROC) curve [7] is another commonly used metric to assess the quality of a classifier. ROC curves plot relationships between the sensitivity and (1 - specificity) while varying the decision boundary. It shows the tradeoff between sensitivity and specificity. The ROC curve of a model is usually quantified by the area under curve (AUC).

### B. Insights of the Problem

In medical data mining, the costs of false negatives are usually much higher than those of false positives [8] because a misclassified malignant tumor is much more fatal than a false alert of healthy issue. Consequently, if a CAD system cannot guarantee perfect sensitivity, inevitably, radiologists have to re-examine the entire believed-to-be negative data out of fear that some positives might slip through the net. In this sense, a CAD system failing to achieve perfect or near-perfect sensitivity cannot really alleviate any of the workloads of medical personnel. Therefore, researchers are motivated to design classifiers that focus on the elimination of all plausible false negatives.

However, minimizing false negatives is not enough because classifiers can easily achieve perfect sensitivity by predicting everything as positive. Therefore, there is a second goal that must be achieved in this task, that is, to maximize the number of true negatives in the predictions. There is a trade-off between these two criteria since the classifiers need to predict more instances as negatives to have a chance to increase the number of true negatives. However, doing so inevitably increases the risk of predicting one or more positive instances as negative. It is this trade-off that makes this problem a difficult one to solve.

### C. A Two-Stage Framework

Acknowledging the two aforementioned trade-off factors to consider in classification, we propose a two-stage framework to handle this problem. The first stage aims at producing a faithful ranking (e.g. using a classification model) of the instances. The second stage then determines a decision threshold that breaks the ranked list of instances into positive part and negative part. Given such a framework, there are two critical issues to be addressed:

1) How to design a learning model that produces a faithful order in which the positive instances are more likely to be ranked above the negative ones.
2) Given the ranked list of instances from 1, how to determine a threshold that has low chance of containing false negatives without having to sacrifice too much specificity.

Note that these two challenges are not necessarily independent since a strategy to identifying the best decision threshold could depend on the quality of ranking. However, tackling them all together makes this a very challenging problem, since then we need to search the model space, parameter space, and threshold space at the same time. Nevertheless, our framework allows the exploitation of the divide-and-conquer principle to tackle two easier tasks independently. We hope by doing so it is more likely to search for a high-quality local-optimal solution.

This study focuses on designing general strategies to tackle the first challenge, which aims at producing a faithful ranking of instances. The second challenge is out of the scope of this work.

## III. A Novel Evaluation Metric: OSPS

We propose a novel evaluation metric called *optimal specificity under perfect sensitivity (OSPS)* that evaluates how well a model ranks *the most difficult positive instance*, which stands for *the positive instance ranked below all other positive ones*. Given an order of instances, OSPS is defined as the maximal specificity attainable with this ranked list of instances. More specifically, to calculate OSPS, the decision threshold is placed at the lowest-ordered positive instance, above which the sensitivity is 100% and below which exists a complete negative set without any false-negative. In other words, OSPS can be illustrated as the proportion of the size of the complete negative set with respect to the total number of negative instances. Consequently, optimizing OSPS is essentially equivalent to optimizing the *upper-bound* of the size of the complete negative set, or equivalently, the optimal amount of effort that can be saved for radiologists. For any given learning model that provides a prediction score or ranking (rather than only a binary decision) for each instance, OSPS can be viewed as the specificity of the negative data in the largest-possible complete negative set. Taking advantage of OSPS, one can now focus on improving the ranking of instances. Mathematically OSPS is defined as

$$\frac{\#(\text{instances}) - \text{idx}(\text{last ranked positive instance})}{\#(\text{negative instances})},$$

where $\#(\cdot)$ counts the size of a set, and $\text{idx}(\cdot)$ denotes the index of an entry in a ranked list.

Several evaluation metrics have been proposed for CAD systems. In addition to ROC, one of the most commonly used metrics for a CAD system is the FROC curve [9]. It modifies the ROC curve by making the y-axis as the sensitivity of the patients rather than instances. Many studies have evaluated

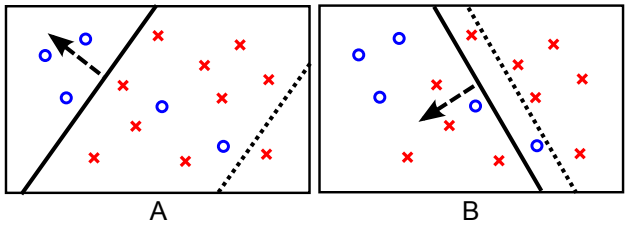


Fig. 1. An example of 15-instance dataset with models A and B. An arrow represents the positive side of a decision boundary. A dotted line represents the decision boundary at which sensitivity is perfect.

their results by area under ROC curve (AUC) [10], [11] within a certain false alert rate [11], [12]. However, doing so raises the difficulty of comparison since people might choose different ranges, and one CAD system might be better for a certain range but not for others.

Note also that improving accuracy, f-score, or AUC does not necessarily guarantee the improvement of OSPS. Like AUC, OSPS is concerned with the ranking of the outputs rather than the absolute classification scores. However, OSPS is a more suitable evaluation metric than AUC in CAD applications. For instance, when considering the design of a CAD system, a model that ranks most of the positive instances on the top of the list but leaves very few at the bottom 1% of the list is not preferred comparing to a model that ranks all positive instances in the middle of the list. As for the former case, the largest possible complete negative set cannot exceed 1% in size; and in the latter case it is possible to find one that is close to 50%. ROC curves fail to reflect such criteria since in the former case, the AUC would be close to perfect (instead of 1%), which is much better than the AUC in the latter case. What makes the difference is that, OSPS precisely targets the threshold at which a CAD system exactly desires; however, unlike OSPS, AUC aimlessly considers every possible decision threshold in its calculation and thus blurs the exact goal of a CAD system.

Another advantage to incorporate OSPS for evaluation is that it is a more suitable criterion for choosing a potentially better model among a set of hypothesized models for CAD systems. For example, Fig. 1 is a sample of 15 instances (5 positives) with two classification models A and B (note that the arrow points to the positive side). Model A apparently has higher accuracy than model B. However, for the purpose of optimizing the size of a complete negative set, if the decision boundary is placed behind the last ranked positive instance for both A and B, it is possible to find that model B is capable of producing a larger complete negative set (5 instances) than that of A (1 instance), as shown by the dotted lines. In this sense, B is better since it has higher potential in terms of reducing the workload of the radiologists (i.e. OSPS value is higher).

According to the definition, OSPS is suitable for application scenarios where the cost of incorrect decision on a positive instance is such tremendous that any false negative is hardly allowed. In some other applications, however, one or few false negatives may be relatively acceptable. We can extend OSPS to support such applications as

$$\frac{\#(\text{instances}) - \text{idx}((K+1)^{th} \text{ last ranked positive inst.})}{\#(\text{negative instances})}, \quad (1)$$

where $K$ denotes the number of false negatives the target application can tolerate.

In general, noises in data can be categorized into two types based on their causes. One type of noises occurs in the feature of the instance, while the other type indicates the incorrectness of the instance's label. The latter type is sometimes incurred by, for example, a mistake made by a human annotator. For the first type of noises, OSPS is a well suited measure for detecting deviated outputs that may be produced when a learning model does not adapt well to noises of this type. For example, let $\boldsymbol{x}$ be a positive instance that is affected by the first type noises. Let A be a model that produces a discordantly low score for $\boldsymbol{x}$, and thus improperly arranges $\boldsymbol{x}$ to be the lowest ranked positive instance. If B is another model that is relatively more insensitive to the noises, and thus grades $\boldsymbol{x}$ into a more proper rank, then B will receive higher OSPS than that from model A.

One may argue that OSPS could be sensitive to the second type of noises. For example, let $\boldsymbol{z}$ be a negative instance that is incorrectly annotated with the positive label. If $\boldsymbol{z}$ is properly assigned a relatively low score by a model C, as it should be, potentially the OSPS of model C could be underestimated. If an application is prone to suffering this type of noises, we can use the extended version of OSPS defined in (1).

## IV. Two Directions to Improve OSPS

This study proposes two instance ranking strategies aiming at improving the OSPS. The first strategy targets particularly at the multi-instance learning tasks, while the second one can be used as a method for more general classification tasks. Both of these two strategies can enhance or cooperate with extant well-developed classification models to produce the ranking scores of instances.

### A. False Positive (FP) Tolerance Method

As described previously, for medical data generally it is the classification of patients (note that each patient has multiple and varied number of instances) instead of instances that is concerned. In this sense, CAD systems generally deal with *multi-instance learning* (MIL) [5], [6] problems: A patient is considered as a set while the ROIs of that patient are considered as instances inside that set. In multi-instance scenarios, a small proportion of false negatives can cause serious damage to the accuracy of the system if they all belong to different sets. On the other hand, increasing the number of correctly identified instances does not necessarily improve the accuracy if the instances are from the same set. Such special features of multi-instance learning bring about different strategies for classification.

Here we consider OSPS in the *patient* level rather than the instance level: To obtain the OSPS for MIL in medical data classification, one has to produce the ranking of the *patients*. A conventional strategy for patient ranking is to first predict the scores of all instances, and then rank the patients based on the highest-scored instance of each. This is a reasonable approach since for MIL, only one positive instance is sufficient to confirm a patient as positive. The OSPS in the patient level

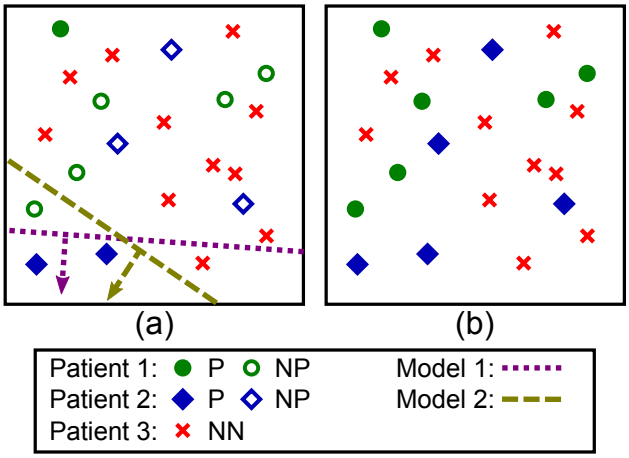


Fig. 2. (a) An example of multi-instance dataset. (b) The dataset after relabeling NP as positives.

is defined as

$$\frac{\#(\text{patients}) - \text{idx}(\text{last ranked positive patient})}{\#(\text{negative patients})}.$$

Given such a strategy, it is not difficult to infer that in order to obtain high OSPS in the patient level, a model has to assign a relatively high score to at least one instance of every positive patient without exception. In this sense, simply applying a conventional classification model for ranking can hardly improve the OSPS since even a single mis-prediction of a positive patient can hurt the overall performance. Therefore we propose a novel false-positive tolerance method to tackle this problem.

In a multi-instance dataset, there are labels at both patient level and instance level. In view of this aspect, we can divide the instances into three categories:

1) Positive instances (P)
2) Negative instances of positive patients (NP)
3) Negative instances of negative patients (NN)

The ranking of the positive patients can be determined by either P or NP, while that of the negative patients can be decided only by NN. One interesting observation is that although raising the ranking of NP for positive patients hurts the instance-level accuracy (since they are truly negative instances), it may in fact improve OSPS. This is because such a strategy might move the positive patient to a higher rank (recall that the rank of a patient is determined by the highest-scored instance, regardless of whether or not it is originally a positive instance). In other words, instances in NP play a very tricky role in MIL in the sense that the misclassifications can improve the patient-level accuracy.

Fig. 2a demonstrates an example (the decision boundaries are represented by dashed lines and the arrows point to the positive side). Model 1 and model 2 both have one false positive instance. However, in model 2, the misclassified instance in NP (the green circle point in the left bottom corner) turns out to suggest the correct identification of positive patient 1.

Based on this observation, one immediate idea would be to re-label NP as positives for classification. However, as can be seen in the later experiments, this is not a promising direction to pursue since labeling all NP as positives produces too much interference for the true positive points, and consequently confuses the classifier. As shown in Fig. 2b, labeling all NP as positives makes it much harder for a model to learn the true positives, and consequently hurts the generalization capability of a classifier.

To deal with such an issue, instead of assigning NP as positives, our strategy is still to assign them as negatives, but to *tolerate* the misclassifications of NP. Compared with the previous proposal, our method does not encourage the misclassifications of NP. Instead we only choose to wink at the false positives of NP. We realized such ideas on SVM and AdaBoost, as described below.

*1) False-positive Tolerance SVM:* We expect to treat misclassified NP less seriously than other types of errors. Based on this idea, we assign different cost values (penalties) for the three categories of data. Normally in medical data, the positive instances are much fewer than the negative ones, and same situation applies to the positive patients and negative ones. That says, there are much fewer P than NP, while NN occupy the majority of the data. For a class-balanced SVM, normally the cost values are assigned based on the distribution of majority and minority classes: The penalties for the misclassifications of the minority class (P) should be larger than that of the majority class (NP and NN). FP tolerance further differentiates NP from NN and assigns different penalties to them.

Built upon standard SVM, *false-positive tolerance SVM* (FPT-SVM) solves the minimization problem

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{l} C_i\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \ldots, l, \end{aligned}$$

where $l$ denotes the size of the training set $S$ and each training instance $\boldsymbol{x}_i$ has a corresponding misclassification cost $C_i$. Let $Q^+$ denote the set of positive patients, which comprises at least one positive instance, and let $B_p$ denote the set of instances belonging to patient $p$. FPT-SVM applies the costs $C_i$ that satisfy the equations

$$\sum_{i\in\text{P}} C_i = \sum_{j\in\text{NP}} C_j + \sum_{k\in\text{NN}} C_k, \tag{2}$$

$$\sum_{i\in B_m\cap\text{P}} C_i = \sum_{j\in B_n\cap\text{P}} C_j, \ \forall m, n \in Q^+, \tag{3}$$

$$C_i = \gamma C_j, \ \forall i \in \text{NP}, \ \forall j \in \text{NN}, \tag{4}$$

where $\gamma$ is a constant between 0 and 1.

In equation (2), the summation of the misclassification costs of all positive instances is equal to the summation of the misclassification costs of all negative instances. To treat all positive patients equally, (3) makes sure that each individual positive patient has equal summation of misclassification costs on the positive instances. In (4), a parameter $\gamma$ is introduced to adjust the ratio of weights between NP and NN. $\gamma$ is a real value between 0 and 1 that makes sure the cost values of NN

are larger than that of NP. A large $\gamma$ does not distinguish NN and NP and a small $\gamma$ tends to ignore the misclassified NP. FPT-SVM can be implemented easily using available SVM packages supporting instance-individual weight assignment. For those that do not support the instance-individual cost weight, one can still achieve FP tolerance by re-sampling the dataset.

*2) Applying FP Tolerance Method to AdaBoost:* The FP tolerance method can also be implemented through adjusting the cost functions of other classification algorithms. Here we apply FP tolerance method on AdaBoost [13]. We assign the importance weights to the base learner in the first iteration as suggested in [14]. In addition, we consider the weight vector $D$ of AdaBoost, which is used to represent a cost distribution of the training instances. For FP tolerance, the elements in $D$ are assigned with $C_i$, which are derived by solving the equations (2)−(4).

*3) Back to the Example in Fig. 2a:* Refer back to the example illustrated in Fig. 2a. Model 1 is a normal model trained without adjusting the costs of NP misclassifications, while model 2 applies false positive tolerance, and the NP misclassifications are assigned with smaller penalties in the learning stage. It turns out that the ranked list of patients produced by model 1 is $(2, 3, 1)$, where positive patient 1 is improperly given a lower score than that of negative patient 3. In contrast, with false positive tolerance, model 2 produces higher score for patient 1, and patient 1 is thus ranked before patient 3. The ranked list becomes $(2, 1, 3)$. Under the measurement of OSPS, model 2 performs better and stands out.

### B. Suspicion Expansion Method

Being able to avoid false negatives is important for not only medical but other types of data such as fraud detection and homeland security. However, data in other domains might not possess the multiple-instance characteristic. Hence, we propose a general idea to improve the OSPS for them.

Because the last-ranked positive point directly affects the OSPS, hard-to-detect positive data can no longer be ignored or treated as noises. Traditionally, a discriminative or generative model employs the strategy of classifying an instance as negative if the probability of it being negative is larger than that of it being positive. However, for OSPS, a false negative can be so severe that we need to adopt a new strategy to enforce stricter criteria about being negative, which is equivalent to having looser criteria for positives. To do so, we propose to generalize the positive class into another class called the *suspicious* class, and constrain the negative class into a more stringent *non-suspicious* class.

Fig. 3 illustrates this idea: The suspicious class contains the original positive class, while the non-suspicious class is a subset of the original negative class. We refer to this process of redefining the labels of data as *suspicion expansion*. We hope by doing so, the instances that are left in the non-suspicious class are those that are highly unlikely to be positive. Therefore this non-suspicious class has a higher chance of being a complete negative set.

One unsolved piece of the puzzle remains: How to expand the positive class into a suspicious one. Considering a case

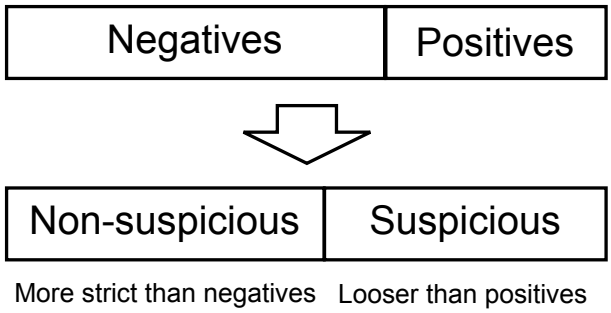


Fig. 3. The suspicious expansion method.

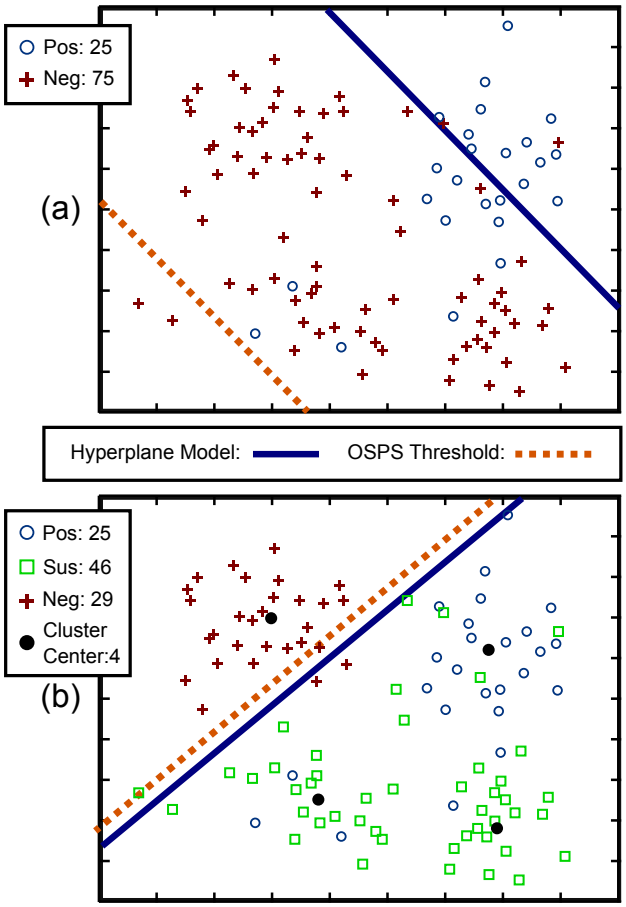


Fig. 4. (a) Training without suspicion expansion. (b) Training after suspicion expansion.

in which one or very few positives are surrounded by many negative points, as shown in the bottom-left part of Fig. 4a. A conventional accuracy-driven classifier would probably treat those positives as noises and still classify all points in that region as negatives. While emphasizing the complete negative set, we can no longer ignore the sparse positive points because although labeling those positives as negatives can improve the generalization capability of a classifier, it can also significantly hurt the OSPS since there are indeed some positives in that region.

Instead, we would like to treat positive instances surrounded by negative ones as the indicator of a suspicious area, and mark all points in that area as suspicious. To be more general, a mass of close instances is labeled as suspicious if at least one positive instance exists within it. To realize this idea, one can first *cluster* the whole dataset into several groups, and identify the suspicious groups as those containing at least one positive instance. Finally we can re-label all points in the suspicious groups as suspicious and the rest as non-suspicious.

Fig. 4 uses a simple classification example to demonstrate this idea. Fig. 4a represents the dataset with original labels (positive and negative) and Fig. 4b represents the re-labeled one after clustering (assuming there are 4 groups). In Fig. 4a, using the linear SVM model for the original labels, we obtain very high accuracy (86%), but the OSPS is low (2%). This happens because of several sparse positive instances located in the left bottom corner. However, after conducting a clustering analysis, we essentially identify four clusters, as shown in Fig. 4b: Three of them are suspicious since they have at least one positive data and one (centered in the upper left part) is non-suspicious. We can then re-label the data as suspicious/non-suspicious: The original 25 positives remained suspicious; 46 out of the 75 negative instances now become suspicious; while only 29 negatives in the upper left cluster remained non-suspicious. Based on the new labels the classifier is capable of obtaining a completely different hyperplane which improves the OSPS to 26%.

## V. Experiments

### A. Dataset Description

Our experiments are performed on the breast cancer dataset, which is a public medical image classification dataset[1] containing patient information. The dataset aims at breast cancer detection on mammogram provided by Siemens Medical Solutions USA. The features are extracted from mammograms for two different views (Mediolateral Oblique and Cranio-Caudal, MLO and CC) of left/right breasts. It is a multi-instance dataset since each patient contains a different number of instances, and the evaluation focuses on the labels of the patients (not the label of the instances). The dataset contains a training set and a test set. The training set consists of 117 dimensions and 102,294 instances of 1712 patients, while the test set contains 94,230 instances of 1602 patients.

### B. Experimental Setup

The two proposed strategies, suspicion expansion and false-positive tolerance, extend the power of single-instance classification algorithms to handle the ranking task of a multi-instance problem. Two classification toolkits are used in our experiments: LIBLINEAR [15] and AdaBoost [13]. Both classifiers can produce real-value outputs as the confidence of an instance being positive. We can then use them to rank the instances and produce the OSPS. The base learner we use with AdaBoost is the classification and regression tree (CART). Note that the parameters of the classifiers are tuned using log-scale grid search to optimize OSPS. For each classifier we compare the performance of four different settings, as shown below:

[1] http://www.sigkdd.org/kdd-cup-2008-breast-cancer

TABLE I. OSPS ON TEST SET

| Classifier: LIBLINEAR | OSPS |
|---|---|
| Original | 12.95% |
| SIL approach | 16.49% |
| FP tolerance method | 20.63% |
| Suspicion expansion method | 23.83% |
| **Classifier: AdaBoost** | **OSPS** |
| Original | 11.68% |
| SIL approach | 8.88% |
| FP tolerance method | 30.77% |
| Suspicion expansion method | 22.70% |
| **Multi-instance Learning** | **OSPS** |
| Diverse Density | 7.14% |
| EM-DD | 10.75% |

1) *Original classifiers*: Conventional classifiers without any modification except parameter tuning. These are treated as the baseline.
2) *Single instance learning (SIL) for multi-instance tasks*: The SIL approach assigns the label of the patient to all its instances. This means that all instances belonging to a positive patient are considered as positive instances, and vice versa for negative ones. This is a simple and typical solution of multi-instance learning.
3) *Suspicion expansion via cluster analysis*: For the suspicion expansion method, we apply K-Means as the clustering algorithm.
4) *False-positive tolerance assignments*: We adjust the cost function in AdaBoost to achieve FP tolerance. For LIBLINEAR, we apply up-sampling on the original dataset.

As the breast cancer dataset by nature is a multi-instance problem, we also compare our strategies to two state-of-the-art MIL algorithms (according to the experimental study in [16]): Diverse Density (DD) [5] and EM-DD [6]. We exploited the package MI-Ensemble[2] [16]. In the training stage, the goal is to find (1) an instance in the feature space that approaches the underlying hidden concept point, and (2) the corresponding feature weights. For DD, we randomly select 10 positive instances in the training set as the starting points, and define the concept point as the one that has the highest diverse density. For EM-DD, the start point is assigned as a randomly selected positive instance. In the test stage, for a test instance, we use its minus weighted L2 distance to the concept point as the confidence score of this instance being positive. Analogously, these scores can then be used to produce OSPS.

### C. Results

In accordance with the evaluation method used in many CAD systems, we compare the OSPS in *patient* level among different strategies. The patients are ranked based on the maximum score of their instances. Evaluation is conducted on the test set. The OSPS for each method on the breast cancer dataset is shown in Table I.

The result shows that for the non-linear classifier AdaBoost, SIL performs even worse than a normal classifier in terms of OSPS. As discussed above, this is mainly because assigning NP as positives confuses the classifier, which in turn hurts its capability to identify the true positive data. Both of

[2] The source code of the package MI-Ensemble is available at http://lamda.nju.edu.cn/code_MIL-Ensemble.ashx

our methods lead to decent improvement for linear SVM and a significant improvement for AdaBoost, Diverse Density and EM-DD.

After analyzing the test set used in the experiment, we found it has relatively inconsistent distribution compared with the training set. It is known that when the test set has slightly different distribution with the training set, as in the case of the breast cancer dataset, a method that tends to overfit cannot generalize well to predict the data in the test set. The results show that, comparing to the competitors, our method is more general and does not overfit the data as seriously. We believe this is because our methods incorporate looser constraints while trying to fit the data, through tolerating FP or generalizing the definition of positives.

In summary, the best result come from the FP tolerance method with AdaBoost, which reaches 30.77% in terms of OSPS, outperforming the best baseline result (16.49%) by 14%, and the best MIL result (10.75%) by 20%.

## VI. Conclusion

One important mission for a CAD system is to produce a complete negative set with no positive instances. Radiologists can then confidently ignore data in this set and focus on the remaining data. The larger the complete negative set, the more effort can be saved. We believe our study has made significant progress on solving this problem, for three reasons.

First, we argue that instead of solving this problem once and for all, we should decompose the problem into two subtasks: The first is to produce a high-quality ranking order that is capable of promoting the rank of the lowest-ranked positive instance. The second is to determine a proper decision threshold to obtain a complete negative set with acceptable size. Under this framework, in the future, researchers can replace the existing ordering and thresholding strategies with better techniques to obtain better results.

Second, we propose an evaluation measure, *optimal specificity under perfect sensitivity*, suitable for medical data classification. We argue that optimizing this metric is equivalent to optimizing the maximal amount of effort that can be saved for radiologists. Another advantage of OSPS is that it allows for comparison among hypothesized models, where models potentially capable of producing larger true negative set, or models insensitive to the noises on features, will be promoted.

Finally, we propose two strategies with different foundations to improve OSPS: (1) The *false positive tolerance* method is designed specifically for multi-instance learning problems, based on assigning not-as-severe penalties to the misclassifications of the negative instances for positive patients. (2) The general-purpose *suspicion expansion* method propagates positive labels to nearby suspicious instances, in order to avoid the misclassification of any positive instance.

Experimental results show the two proposed approaches not only improve OSPS significantly (as much as 1.63 times better than a conventional classifier), but also avoid the potential overfitting problem. Our methods are not overwhelmed by the relatively inconsistent distribution between the training set and the test set.

The evaluation criterion OSPS and related learning strategies proposed in this work are not limited to medical data. Generalized application to other scenarios where false negatives imposed extremely high costs (e.g. cyber-attack or intrusion detection, or detection of catastrophic failures of public facilities) might be beneficial as well.

One main future work of us is to solve the second part of the puzzle, which is to approach the ideal decision threshold that produces perfect sensitivity while still being able to filter significant amount of negative instances.

## References


[1] G. Fung and J. Stoeckel, "Svm feature selection for classification of spect images of alzheimer's disease using spatial information," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 243–258, 2007.

[2] M. A. Jaffar, A. Hussain, and A. M. Mirza, "Fuzzy entropy based optimization of clusters for the segmentation of lungs in ct scanned images," *Knowledge and Information Systems*, vol. 24, no. 1, pp. 91–111, 2010.

[3] R. M. Rangayyan, F. J. Ayres, and J. E. L. Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," *Journal of the Franklin Institute-Engineering and Applied Mathematics*, vol. 344, no. 3-4, pp. 312–348, 2007.

[4] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1-2, pp. 1–24, 2002.

[5] O. Maron and T. Lozano-Prez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*. MIT Press, 1998, pp. 570–576.

[6] Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 1073–1080.

[7] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006.

[8] H. M. Zhao, "Instance weighting versus threshold adjusting for cost-sensitive classification," *Knowledge and Information Systems*, vol. 15, no. 3, pp. 321–334, 2008.

[9] D. P. Chakraborty and I. Brikman, "Free-response receiver operating characteristic (froc) analysis in medical imaging," *Medical Imaging Iv : Pacs System Design and Evaluation, Parts 1 and 2*, vol. 1234, pp. 517–526, 1990.

[10] G. Fung, M. Dundar, B. Krishnapuram, and R. Rao, "Multiple instance learning for computer aided diagnosis," *Advances in neural information processing systems*, vol. 19, p. 425, 2007.

[11] H. Soltanian-Zadeh, F. Rafiee-Rad, and S. D. Pourabdollah-Nejad, "Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms," *Pattern Recognition*, vol. 37, no. 10, pp. 1973–1986, 2004.

[12] S. Yu, K. Li, and Y. Huang, "Detection of microcalcifications in digital mammograms using wavelet filter and markov random field model," *Computerized Medical Imaging and Graphics*, vol. 30, no. 3, pp. 163–173, 2006.

[13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[14] Y. Sun, M. Kamel, A. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[15] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[16] Z.-H. Zhou and M.-L. Zhang, "Ensembles of multi-instance learners," in *Proceedings of the 14th European Conference on Machine Learning (ECML'03)*. Springer, 2003, pp. 492–502.