Abnormality Detection in Retinal Images

Yu Xiaoxue[1], Wynne Hsu[1], W. S. Lee[1], Tomás Lozano-Pérez[2]
[1]Singapore-MIT Alliance, CS Program, National University of Singapore, Singapore117543
[2]NE43-719 200 Technology Square Cambridge, MA 02139-4307, U.S.A.

***Abstract*— The implementation of data mining techniques in the medical area has generated great interest because of its potential for more efficient, economic and robust performance when compared to physicians. In this paper, we focus on the implementation of Multiple-Instance Learning (MIL) in the area of medical image mining, particularly to hard exudates detection in retinal images from diabetic patients. Our proposed approach deals with the highly noisy images that are common in the medical area, improving the detection specificity while keeping the sensitivity as high as possible. We have also investigated the effect of feature selection on system performance. We describe how we implement the idea of MIL on the problem of retinal image mining, discuss the issues that are characteristic of retinal images as well as issues common to other medical image mining problems, and report the results of initial experiments.**

***Index Terms*— Data mining, abnormality detection, multiple-instance learning, medical image mining.**

I. Introduction

Abnormality detection in images is predicted to play an important role in many real-life applications. One key example is the screening of medical images. A fast, accurate, and reliable method for abnormality detection in images will help greatly in improving the health-care screening process.

Existing efforts in abnormality detection have largely been focused on relational databases [5]. They can be categorized into two main approaches. In the first approach, a standard of what are the norms is first established. Significant deviations from the established standards is regarded as abnormal. Techniques such as statistical deviation analysis and clustering [7] fall into this category. A different approach is to learn the characteristics of the abnormalities through supervised training. Based on the learned characteristics, detection is performed by checking whether any of these learned characteristics is present in the data. Techniques in this approach includes: case-based reasoning and classification using neural networks, decision trees, etc. [7].

We believe that the second type of approach, based on building a classifier for the abnormalities of interest is better suited to detecting abnormalities in medical images, since these abnormalities are more regular than, say, fraudulent ATM transactions. By exploiting what we know of the regularities in both normal and abnormal images we should be able to achieve a better combination of sensitivity and specificity.

Unfortunately, applying supervised-learning based approaches to detecting abnormalities in images is not straightforward. Supervised learning assumes that the training samples are classified by an expert as either normal or abnormal. However, in typical applications, such as medical image analysis, the training images only have vague class labels (normal, abnormal), yet without information from the human experts as to what aspect of the image justifies the label.

To learn the characteristics of abnormalities, we adapt the multiple-instance learning framework of [2]. We use the idea of multiple-instance learning (MIL) because of the incomplete labelling information that we have about the training images. Namely, the training images only have class labels (normal, abnormal), yet without information from the human experts as to what part of the image caused the abnormal labels to be applied. Hence, the MIL framework is an appropriate approach to the problem. To the best of our knowledge, this is the first attempt at using multiple-instance learning for learning abnormalities in medical images.

In this paper we present a framework for abnormality detection in images. The framework consists of three main components: extraction of relevant image features, the discovery of abnormality characteristics and dealing with errors in the training data.

To extract relevant image features, we use the algorithm described in [10] that automatically discovers relative invariant relationships among objects in a set of images. Here, we show how the algorithm can be utilized to suggest suitable image features to be used for subsequent abnormality detection process.

The learning process is complicated by the fact that it is important to maintain high sensitivity, particularly for medical image screening. In other words, we cannot afford to miss out any abnormality if one truly exists. A high sensitivity multiple-instance learning strategy is proposed. Experiments on real-life data set of diabetic retinal screening images show that we are able to achieve some improvement upon a previous system [6] without any additional time cost.

Another important aspect of abnormality detection in medical images is that a medical image data is typically very noisy, which is a characteristic of all medical data.[19][12] To obtain meaningful detection accuracy, we need, in particular, to deal with incorrectly labelled training data. Our approach to this within the context of the MIL frameworks is described below.

The outline of the paper is as follows. In Section 2, we discuss related work in the areas of abnormality detection and medical image mining. Section 3 gives our approach to solve this problem, including image pre-processing, feature generation and learning. Section 4 shows the experiment results. Finally, we conclude in Section 5.

Ms. Yu: smap1063@nus.edu.sg
Dr. Hsu: whsu@comp.nus.edu.sg
Dr. Lee: leews@comp.nus.edu.sg
Prof. Lozano-Pérez: tlp@mit.edu

II. Related Work

Here we define "Abnormality Detection" as detecting rare abnormal objects in a data set, via modelling the rare objects themselves, instead of the normal objects. For example, if we want to detect a certain fraudulent activity in the usage of credit cards, we would do it by learning a definition of such fraudulent activity, followed by detecting any such activities in the customer usage data according to this definition. This is quite different from approach of learning the characteristics of normal objects first, and then classify sufficiently different objects as abnormal ones. Which approach is preferable will depend on the application domain. We believe that characterizing the abnormal objects directly is appropriate for the medical image domain.

Recently, there has been an increased interests in medical image abnormality detection. Parr, *et al.*, applied model based classification to detect anatomically different types of linear structures in digital mammogram, in order to enable accurate detection of abnormal line patterns.[16] El-Baz and his colleagues proposed a method based on analysis of the distribution of edges in local polar co-ordinates to detect lung abnormalities in 3-D chest spiral CT scans [3].

Such detection is a challenging task due to the similarity between the real abnormalities and other normal patterns in the images. Consequently most algorithms produce a large amounts of false positives.[16] Here, we represent the medical image abnormality detection problem as a multiple-instance learning (MIL) problem,[2][14] in order to reduce the large amount of false positives while maintaining the true positives detection accuracy. [17], [21] and [11], have applied the multiple-instance learning framework to images where the main target is to detect the existence of some pre-defined objects in a series of images. Their problem is slightly easier since the demands on sensitivity are very different.

III. Our Proposed Approach

A. Overview

Figure 1 gives an overview of the major steps involved in the detection of abnormality in images. Initially, a subset of the images are processed to obtain the relevant features for learning the characteristics of the abnormalities in the images. [10] makes the observation that in images, more often than not, it is the relative relationships among the image features that are meaningful for interpretation. Based on this observation, we employ the ViewpointMiner[10] to first discover the significant patterns among features. These patterns serve as important hints as to the type of image features to be extracted for the subsequent learning process.

Once the training images have been processed to extract the relevant image features, we transform the problem of abnormality detection in images to a multiple-instance learning problem. We introduce the notion of "Fake True Bag" to deal with outliers among the training images. This allows us to reduce the large number of false positives while maintaining a high-level of true positive detection in the presence of noise. When the learning is completed, the results are then used to detect the presence of abnormality in the new set of images. Details of each step are provided in the following subsections.

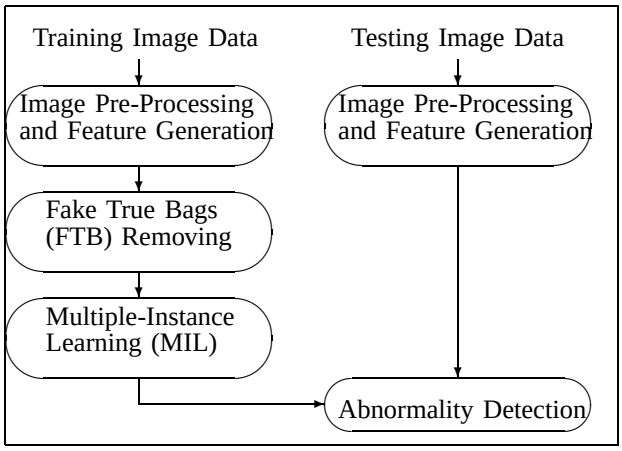

Fig. 1. Image Abnormality Detection Using Multiple-Instance Learning

B. Image Pre-Processing Feature Selection

Most high-level image feature extraction requires some form of human intervention. We do not have a good mapping that maps the low-level image features (color, texture, etc.) to the high-level image features (lesions, blood vessels etc.). As a result, most image feature extraction processes are domain-specific and highly specialized. Therefore, it is important to give the background of the type of medical images we are dealing with in this paper before we proceed to describe the image feature extraction process.

In this paper, we focus on retinal images. In particular, we attempt to identify abnormalities relating to diabetic retinopathy. One of the symptom of diabetic retinopathy is the presence of exudates. Exudates show up as random bright patches around the inter-vascular region. They vary in shapes and sizes (see Figure 2). A number of methods have been proposed to detect the presence of exudates through image processing techniques. [13] shows that using the features such as size, shape and texture in isolation is insufficient to detect hard exudates accurately. If the background color of a retinal image is sufficiently uniform, a simple and effective method to separate exudates from the background is to select proper thresholds [20]. [9] has developed a system for detecting the presence of exudates based on clustering and classification techniques.

Among the features extracted by [9] are:

- **Size**: The size of a suspected exudate.
- **Average Intensity**: The average intensity of a suspected exudate.
- **Average Contrast**: The average intensity of the suspected exudate with respect to the average intensity of its surrounding region.
- **Optic Cup-Disc Ratios**: Figure 3 shows the optic disc (the bright area within the gray oval). The optic cup is the brightest region within the disc. According to domain knowledge, the ratio of the cup-disc is typically 0.3. If the ratio is too large, it indicates that a detected bright spot is likely to be a true positive.

Fig. 2. Original retinal image: The hard exudates (the bright patches) will be two spots on the processed image

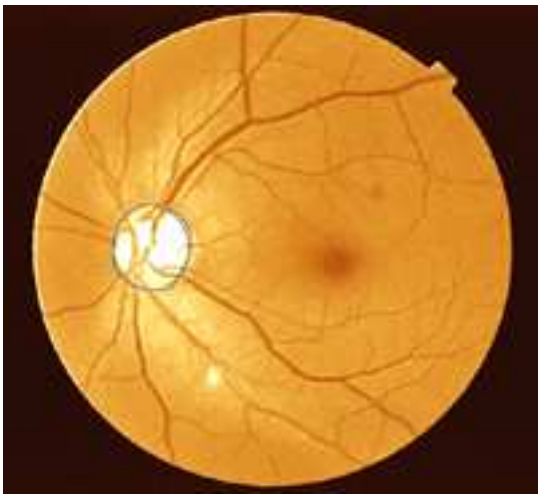

Fig. 3. Detect Optic Disc in a Retinal Image

While these features are able to maintain a 100% true positive detection rate, their false positives detection rate is very high(about 76.52% of images detected as "with exudates" are actually without any exudates).

Here, besides the features mentioned above, we employ the ViewpointMiner [10] to discover significant relative relationships in the set of retinal images. The essential idea behind ViewpointMiner is to iteratively generate k-objects patterns from $k-1$-objects patterns such that certain fixed distance or orientation relationships are maintained. Here, we fix one of the objects to be the optic-disc and discover that the exudates exhibit interesting distribution around the optic disc (see Figure 4). The darker the region is, the more patterns are instances are foud within this region.

Clearly, the relative distance between a suspected exudates and the optic disc is a useful feature for detecting false positives. We extract this feature as follows.

Optic Disc Distance: the optic disc distance of each suspected exudate is the ratio of the distance between the suspected exudate and the center of the optic disc to the distance between

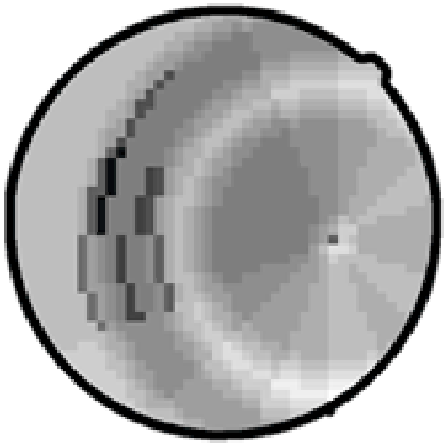

Fig. 4. Exudates Distribution According to the Optic Disc

the center of the optic disc to the periphery of the retina.

To generate the feature values above, we take the processed images of ADRIS exudates detection[6] and the original retinal images as our input data. Each region on the processed images will be recorded as a potential exudate, and we apply some simple image growing methods to combine several regions that are very close to each other into a bigger region. Such a combination was implemented based on data from the original image, i.e., if these regions correspond to several bright patches close to each other in the original retinal image, we take them as several parts of one unique region which was split during the processing in the ADRIS system. After detecting the positions of the regions in the images, we can generate the corresponding feature values of each region.

Once we have transformed the retinal images into a table of extracted features, we proceed to the learning phase.

C. Multiple-Instance Learning

One characteristics of medical images is that as long as there exists **one** abnormality in an image, that image is classified as abnormal. All existing retinal image abnormality detection algorithms reviewed in this paper make the assumption that each potential abnormality region in the training data set has been correctly labelled by human experts or according to some pre-defined rules. However, this is certainly not true in practice. Medical professionals only label an **image** as normal or abnormal. They do not look for the presence of all abnormalities in an image before concluding that an image is abnormal.

To model such a scenario, we represent our problem as a multiple-instance learning problem. The multiple-instance learning problem can be easily understood by the following example. Suppose there is a lock that we want to open, and there are some key chains available. We know which key chains contain the key that can open the lock. Without testing all the keys in those key chains, can we find a pattern for the desired key(s)? In other words, if we are given bags of examples and we know the label of each bag (positive or negative), can we infer the defining characteristics of the positive examples.

The multiple-instance learning problem was first introduced in [2]. Subsequent works [14] [17] [21] [22] extend the framework and apply it to other applications. Here, we map our retinal image abnormality detection problem to a multiple-instance learning problem as follows:

- **Instance**:Each instance consists of the extracted features of a suspected exudate region (see Section 3.2).
- **Bag**:Each bag refers to a retinal image, and it contains all the instances within this image.
- **True Bag**:A bag is labelled as "true", i.e., "with exudates", if and only if at least one of the instances within this bag is a true positive.
- **False Bag**:A bag is labelled as "false", i.e., "without exudates", if and only if none of the instances within this bag is true positive.

For ease of reference, from now on, we will call those instances from true bags *True Bag Instances*, while those instances from false bags will be called as *False Bag Instances*. Note that not all instances from true bags are true positives, that is, actual exudates. To solve the multiple-instance learning problem, we implement two different algorithms, the Diverse Density Algorithm (DD) and the Axis-Parallel Rectangles Method (APR).

1) Diverse Density Algorithm: The basic idea behind the Diverse Density (DD) algorithm introduced in [14] is to try to find certain point in the feature space whose ratio of the density of true positives to that of false positives is highest.

For our retinal images, DD assumes that there exists one "typical true instance". In the area near this instance, one can expect that the density of true bag instances to be high, while the density of the false bag instances to be low. We define "Diverse Density" as the ratio of the true instance density to the false instance density. With this, we can find the typical true instance by maximizing the diverse density in the feature space.

Let us denote the i^{th} true bag as B_i^+, the j^{th} instance in that bag as B_{ij}^+, and the value of the k^{th} feature of that instance as B_{ijk}^+. Likewise we denote B_i^-, B_{ij}^- and B_{ijk}^-. Assume the typical true instance we are looking for is a single point t in the feature space, and with the additional assumption that the bags are conditionally independent given the target t, we represent the problem as maximizing the following value:

$$V(t) = \arg\max_x \prod_i \Pr(x = t \mid B_i^+) \prod_i \Pr(x = t \mid B_i^-) \quad (1)$$

And for a bag B_i(either B_i^+ or B_i^-), using a noisy-or model, we can get the probability that not all instances miss the target is:

$$\Pr(x = t \mid B_i) = \prod_j (1 - \Pr(x = t \mid B_{ij})) \quad (2)$$

We model the causal probability of an individual instance on a potential target as related to the distance between them:

$$\Pr(x = t \mid B_{ij}) = \exp(- \parallel B_{ij} - x \parallel^2) \quad (3)$$

Finally, the distance between two points in the feature space is calculated by:

$$\parallel B_{ij} - x \parallel^2 = \sum_k s_k^2 (B_{ijk} - x_k)^2 \quad (4)$$

while s_k are the pre-defined weighting factors.

To search the feature space, we implement a gradient descent search method (see Algo GradientDD) with some initializing strategy. For the gradient search method, the choice of the starting point is very important. By intuition, the instances from the true bags should be some good starting points. Hence we use gradient search to find one "hypothesis" starting from each instance of the "starting true bags", based on the training set. These hypotheses will be validated upon a validation set, to choose the best one as the final target.

The gradient descent search algorithm requires a pre-defined step length r to do the search. For each step, we calculate the k-D gradient of $V(x)$, $\nabla V(x)$, and use this gradient to get to the next point. The gradient search will end till the difference between the latest two *V* values is small enough (We use the parameter *MinDiff* to tune the minimum distance in our algorithm).

Algo GradientDD

```
1  Randomly choose N starting bags B_1, B_2, ..., B_N
2  for each B_i
3       {for each instance I_j within B_i
4            {x = I_j
5            v = V(x) (based on the training set)
6            x' = x + ∇ V(x)
7            while (| V(x') − v |> (MinDiff × v))
8                 {x = x'
9                 v = V(x)
10                x' = x + ∇ V(x) ×r }
11           h_ij = x' }
12      }
13 bs = 0
14 p = 0
15 q = 0
16 for each h_ij
17      {Test h_ij on the validation set to get the
             sensitivity sen and specificity spec
18      if (sen > MinSen)
19           {if (spec > bs)
20                {bs = spec
21                p = i
22                q = j }
23           }
24      }
25 return(h_pq)
```

In order to define the "best hypothesis" during the validation step, we use the parameter of *MinSen*. Only those hypotheses with validated sensitivity (based on the validation set) higher than the *MinSen* value will be considered, and then the one with the highest validated specificity will be chosen as the final target.

Initial investigations show that the Diverse Density algorithm does not always give robust performance, especially when we require the sensitivity value beyond 95%. This is because the assumption of one "typical true instance" sometimes

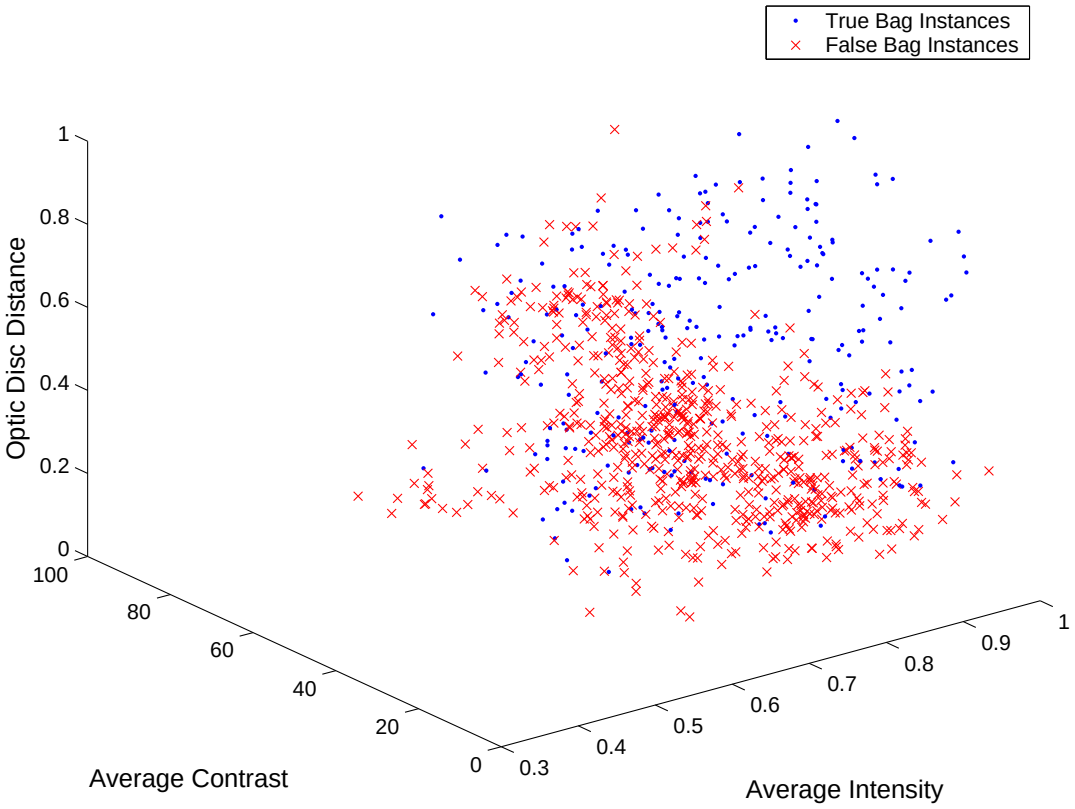

Fig. 5. Distribution of Instances in Feature Space

doesn't hold in the retinal images. Figure 5 shows the region occupied by the true instances. It is clear that the true instances occupies a rather large area, and the distribution within this area is almost flat. [1] So, the actual choice of a typical point in the feature space is rather arbitrary. Hence we also tried another MIL method, i.e., the Axis-Parallel Rectangles method[2] to solve our problem.

2) Axis-Parallel Rectangles Method: The Axis-Parallel Rectangles method (APR) for MIL was first introduced in [2]. The basic idea is to draw an axis-parallel rectangle in the feature space to cover at least one instance from all the true positive bags, and to exclude as many false positive instances as possible. Once we have constructed such a rectangle, it can then be used to judge whether a incoming bag is true positive or not.

Several algorithms to learn an axis-parallel rectangle for MIL problems have been developed in [2]. In this paper, we first modify the axis-parallel rectangle framework to fit our retinal image problem, and then implement a greedy shrinking algorithm to learn the rectangle. Before we describe our algorithm, we first introduce some definitions.

- **Allpos APR**: In the k-D feature space, we can draw a minimum sized axis-parallel hyper-rectangle that covers all instances from the true bags. We call this APR Allpos APR.
- **Shrinking Cost**: When we attempt to shrink an APR, the cost of each step s is calculated by:

$$C(s) = \frac{NumofExcludedTrueBagInstances}{NumofExcludedFalseBagInstances} \quad (5)$$

  Similarly, the value of 1/C can be defined as the **shrinking gain**.
- **Minimax APR**]: We can shrink the **Allpos APR** to get a minimum (maybe locally) APR that contains at least one instance of all true bags, which we call the Minimax APR. Here the "minimum" means along all the axis, any further shrinking will either exclude one true bag totally, or make the **shrinking cost** exceed a certain threshold.

Our target is to find one Minimax APR[2] in the feature space during the training. When a new image is obtained, we can decide whether it has exudates by judging whether each of its candidate exudate regions falls into the Minimax APR or not. Figure 6 shows an example of the learned APR on a set of real-life retinal images.

To get a Minimax APR, we applied an outside-in greedy shrinking strategy, shown as Algo APRShrink.

In Algo APRShrink, the *MaxCost* means the maximum cost we can allow to do the shrinking. If one shrinking step will cause the exclusion of one whole true positive image, we will set the shrinking cost as $MaxCost$ also. Let m denotes the number of instances from the true positive bags, the time

[1] Here we only show a 3D projection of the feature space, while the real feature space is at least 4D.

[2] In general, there is no unique Minimax APR. In a high-dimensional space there will generally be many, which will generalize differently. The algorithm we used greedily picks one such.

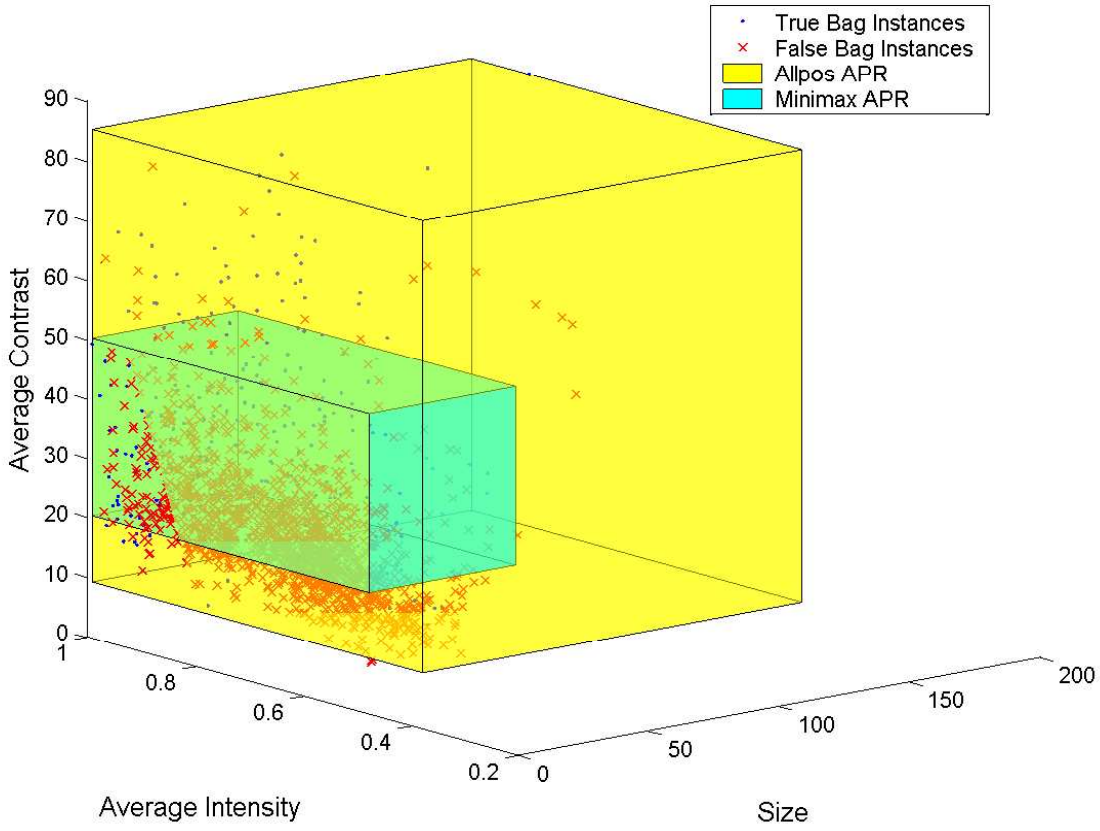

Fig. 6. Learned APR of Retinal Images

complexity of line 4 to 19 is $\Theta(k \times m)$, and if we maintain a matrix to record whether one instance is excluded or not, each shrinking operation will cost $\Theta(k)$ time. So the time complexity of this algorithm is determined by the number of shrinking operations, which is determined by the step length of the shrinking.

Another way to do the APR shrinking is to shrink the APR by excluding one instance at a time, instead of with a certain step length. We adopted a fixed step-length strategy for simplicity and efficiency. Our investigation shows that even if we chose the step-length to be only 0.1% of the range of each axis, the Minimax APR can be obtained within seconds and produce a Minimax APR that obtains excellent results. This is because line 1 has reduced the whole feature space by about 1/3, as shown in Figure 6. Furthermore, many of the axis directions are quickly labelled as "unshrinkable" because the true instances are almost at the edges of the Allpos APR. All in all, we only need to deal with 4 to 6 directions (note that each axis has two directions) on average which makes the algorithm highly efficient.

3) Dealing with Noisy Data: As we have mentioned before, the retinal image data is very noisy. The noise comes in two forms: One is from the many false instances in both true bags and false bags. In addition, it is possible that due to human error, some bags have been labelled wrongly as "true" bags. The existence of such *fake true bags* results in bad performances of the learning algorithms, especially the APR method. This is because such fake true bags tend to be very far from other true bags in the feature space. In other words, they serve to enlarge the generated APR greatly and hence reduce the effectiveness of the algorithm. The DD algorithms seem not too sensitive to the existence of such fake true bags, since we start the search from different starting points such that the local optimal introduced by the fake true bags could be removed during the validation step. To identify the fake true bags in the training data set, we define them as follows:

Fake True Bag (FTB): If one true bag can be removed from the Allpos APR by shrinking the APR along one axis direction, while no other true bag will be affected, and the shrinking efforts till the next true bag can exclude enough false instances (the **shrinking gain** g is beyond a given threshold), we call this true bag an Fake True Bag.

This motivated the step of FTB Removal before the APR shrinking, as a technique of *Outlier Rejection* (see Algo FTBRemoval).

In Algo FTBRemoval, *MinGain* is the pre-defined shrinking gain threshold. The detection and removal of FTBs does not increase the time complexity of the learning method we employed.

Finally, in a mature APR system, the processing of the images should include (1)Image pre-processing and feature

Algo APRShrink.

```
1  draw the Allpos APR R for the training set
2  for (i = 0; i < k; i++)
3       set axis i as "shrinkable"
4  while there exists axis i that is "shrinkable"
5       {for (i = 0; i < k; i++)
6            c[i] = 0
7       for (i = 0; i < k; i++)
8            {if (c[i]=0)
9                 {step s = shrinking R
                  along ax i by one unit
10                c[i] = C(s)
11                if (C[i] > MaxCost)
12                     set axis i as "unshrinkable"
13                }
14           }
15      lc = min(c)
16      li = the one in c[i] corresponding to lc
17      if (lc < MaxCost)
18           {step s = shrinking R along ax li by
                one unit
19           R ← s(R)}
20      }
21 return(R)
```

Algo FTBRemoval.

```
1  draw the Allpos APR R for the training set
2  for (i = 0; i < k; i++)
3     g[i] = 0
4  for (i = 0; i < k; i++)
5     {step s = shrinking R along axis i till
            the second true positive bag
            is going to be excluded
6     if no other true positive bag is
         excluded during step s
7        g[i] = 1/C(s)
8     if (g[i] > MinGain)
9        R ← s(R)
10    }
11 return R
```

generation; (2)FTB Removal; (3)Using APRShrink to find one Minimax APR and (4)Detecting abnormalities in the new images using the learned Minimax APR.

The experiment results of the two algorithms above will be shown in section 4.

IV. Experiment Results

In this section we present the results of the experiments to evaluate our retinal image abnormality detection system. The experiments are carried out on a Pentium 4, 1.6 GHz processor with 256MB memory running Windows XP. All the algorithms are implemented using C++.

A set of 562 real-life retinal images are obtained from primary health-care clinics in Singapore. The quality of these images varies greatly. Some of the images are blurred and over-exposed while others are clear and distinct. Among these 562 images, 132 images have suspected exudates of which 31 are labelled by the human physicians as being "abnormal". These 132 images are used as the input data to our system, for training and testing.

We run our retinal image abnormality detection system by using a 10-fold cross validation method. For the DD algorithm, the data set is partitioned equally into 10 parts of which 6 parts are used as training set, 3 as validation set and the remaining 1 part is used as the testing set. While in other learning methods, we simply use 9 folds as the training set and the remaining 1 as the testing set. Within each fold, there are 3 to 4 true positive images (actually having exudates) and 13 to 14 false positive images. The process is repeated 10 times and the average accuracy is obtained. Two measures are used in our experiments: (1) sensitivity which is defined as the percentage of true positive images that are correctly detected, and (2) specificity which is defined as the percentage of detected negative images within the true negative ones. The calculation of sensitivity and specificity are based on the definition of *confusion matrix* in [4]. Note that both sensitivity and specificity are calculated within the output images from the original ADRIS system labelled as "with exudates". The ADRIS system obtained this image set by processing the original retinal images from the polyclinic and achieved the first-step result, which is 100% sensitivity and almost 25% specificity.

A. Effect of FTB Removal upon APR Method

Here we test the effectiveness of using FTBRemoval to deal with outliers in the data. Figure 7 shows the results.

For the APR shrinking, we chose the $MasCost$ to be

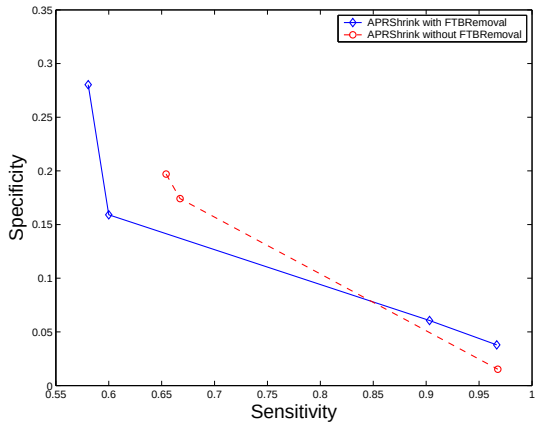

Fig. 7. Effect of FTB Removal in APR Shrinking Methods

100, and a step of excluding one whole true positive image from the APR during training will be considered as with shrinking cost of $MaxCost$. The parameters we simulated during this experiment include the step lengths within the APR shrinking (along different axis and directions) and the "exudate threshold". The exudate threshold is represented as the number

of axes that one instance must satisfy during testing for it to be detected as a real exudate. To be more specific, if the exudate threshold is M, and one instance falls into the learned Minimax APR along no less than M axis, it will be judged as a true positive instance. Such a threshold roughly measures the probability of one instance being a true positive one. For both of the curves shown in Figure 7, the leftmost data points were obtained while "exudate threshold" was 6, and the other(s) were obtained while "exudate threshold" was 5. We have 6 features altogether for each instance. When the "exudate threshold" is fixed, we can change the shrinking step lengths along the axis to get several data points.

In our experiment, we chose the thresholds such like $MinGain$ such that the FTB Removal only removes one bag from the true positive images, and such a removal is remained in the sensitivity result, i.e., we still consider this removed true positive image as one of the wrongly excluded true positive bag of our system. From the curves, we can observe that when the sensitivity remains quite high (beyond 80%) with the removal of the FTB, we can achieve a higher specificity at the same sensitivity than without FTB Removal. And at the point of sensitivity at 96.67%, FTB removal could get about 2% higher specificity. So if we want to keep sensitivity very high, the APR shrinking with FTB Removal is preferred.

B. Diverse Density Search vs. APR Shrinking

Also we tested the different MIL algorithms, i.e., the Diverse Density Algorithm and the APR Shrinking algorithm. Figure 8 gives the comparison of Diverse Density Search method, APR Shrinking with FTB removal and APR Shrinking without FTB removal.

For the gradient DD search, we use 6 true bags as the

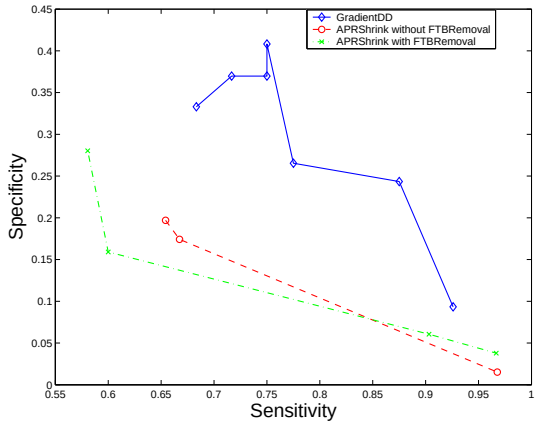

Fig. 8. Comparison between Gradient DD Search and APR Shrinking

starting bags on each run, and the stopping criteria of the gradient search is that the difference between the two latest *V* values (*V(x)* and *V(x')*) is within 0.1% of *V(x)*. Hence we can tune the parameter *MinSen* to get the curve of specificity vs. sensitivity. The GradientDD curve shows that the part with sensitivity below 80% is rather unstable, but since in medical implementations such a low sensitivity case is not quite preferred, we only focus on the part with sensitivity beyond 80%. In this part, the gradient DD search generally beats the APR shrinking. Another point here is that the gradient DD search cannot get a stable result with sensitivity value higher than 95% in our experiment, while the APR shrinking methods can do a little bit better.

C. Comparison between MIL Methods and SVM

To compare the MIL methods with non-MIL methods, we also tested SVM*light*[3] on our experiment data set. Figure 9 shows the comparison between gradient DD search and SVM*light*.

To do the training using SVM*light*, we tuned the parameter

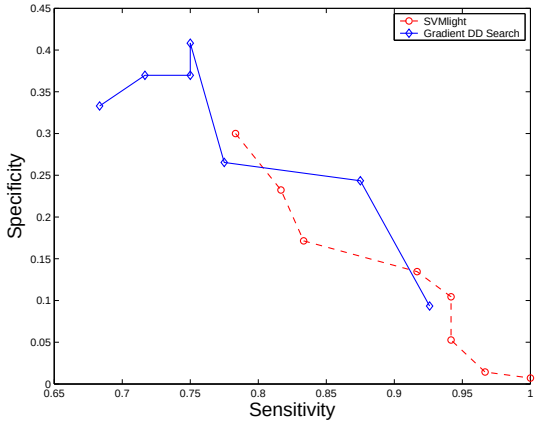

Fig. 9. Comparison between Gradient DD Search and SVM*light*

cost factor to change the weights between positive and negative instances. More details about induction by SVM*light* and the *cost factor* could be found in [15]. In our experiment, we got the specificity vs. sensitivity curve by varying the *cost factor* from 20 to 45. We can observe that actually GradientDD and SVM*light* perform similarly. Because of this we are currently unable to conclude whether MIL methods outperform non-MIL methods for this problem.

V. Conclusion

Multiple-Instance Learning has been shown to be an effective way to solve learning problems where the training examples are ambiguous, i.e., a single example object may have many alternative instances, yet only some of these instances are responsible for the observed classification of this object. For the first time, we tried to implement this method in the area of medical image mining, in particular, as an effective solution to abnormality detection in retinal images. Though MIL has been used to solve some other image mining problems before, there are some unique characteristics of medical images that require extra effort and some new techniques, like the extremely noisy data and the strict requirement of sensitivity. We implemented two approaches to MIL, Diverse Density and APR, to solve our retinal image mining problem. The experiment results suggest that some MIL methods, like

[3]SVM*light* is an implementation of Support Vector Machines (SVMs) in C. The source code, binaries and manuals could be found on the following website: http://svmlight.joachims.org/.

the gradient DD search, is somewhat useful in solving this problem. However, it is not clear from the experimental results whether the MIL methods that were used outperform non-MIL methods for this task. Future work includes exploring other feature representation and other MIL methods for this problem.

REFERENCES

[1] Andrews S., Hofmann T. Tsochantaridis I.: Multiple Instance Learning with Generalized Support Vector Machines. *AAAI/IAAI'2002*: 943-944.

[2] Dietterich T.G., Lathrop R.H. Lozano-Pérez T.: Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence*, 89(1-2):31-71, 1997.

[3] El-Baz A., Farag A.A., Falk R. Rocca R.L.: Detection, Visualization, and Identification of Lung Abnormalities in Chest Spiral CT Scans: Phase I. Technical Report, CVIP Laboratory, University of Lousiville, July 2002. (TR-CVIP-7-02)

[4] Fielding A.H. Bell J.F.: A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. *Environmental Conservation* Vol.24, pp.38-49, 1997.

[5] Galván J.R., Elices A., Muñoz A., Czernichow T. Sanz-Bobi M.A.: System for Detection of Abnormalities and Fraud in Customer Consumption. *12th Conference on the Electric Power Supply Industry*. Pattaya, Thailand. Nov 2-6, 1998.

[6] Goh K.G., Lee M.L., Hsu W. Wang H.: ADRIS: An Automatic Diabetic Retinal Image Screening System. *Medical Data Mining and Knowledge Discovery*, Springer-Verlag, 2000.

[7] Han J. Kamber M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[8] Hsu W., Lee M.L., Liu B. W.L. Tok: Exploration Mining in Diabetic Patients Databases: Findings and Conclusions. *Knowledge Discovery and Data Mining*, pp.430-436, 2000.

[9] Hsu W., Pallawala P.M.D.S., Lee M.L. Kah-Guan A.E.: The Role of Domain Knowledge in the Detection of Retinal Hard Exudates. *IEEE Computer Vision and Pattern Recognition*, Hawaii, Dec 2001.

[10] Hsu W., Dai J. Lee M.L.: Mining Viewpoint Patterns in Image Databases. *SIGKDD2003*

[11] Huang X., Chen S.C., Shyu M.L. Zhang C.: User Concept Pattern Discovery Using Relevance Feedback and Multiple Instance Learning for Content-Based Image Retrieval. *The 3rd International Workshop on Multimedia Data Mining (MDM/KDD'2002), in conjunction with the 8th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pp.100-108, Edmonton, Alberta, Canada. July 23, 2002.

[12] Huyn N.: Data Analysis and Mining in the Life Sciences. SIGMOD Record 30(3):76-85, 2001.

[13] Leistritz L. Schweitzer D.: Automated Detection and Quantification of Exudates in Retinal Images. *SPIE*, Vol. 2298, 690-696, 1994

[14] Maron O. Lozano-Pérez T.: A Framework for Multiple-Instance Learning. *Advances in Neural Information Processing Systems*, MIT Press, 1998.

[15] Morik K., Brockhausen P. Joachims T.: Combining Statistical Learning with a Knowledge-Based Approach - A Case Study in Intensive Care Monitoring. *Proc. 16th Int'l Conf. on Machine Learning (ICML-99)*, 1999.

[16] Parr T.C., Astley S.M., Taylor C.J. Boggis C.R.: Model-based classification of linear structures in digital mammograms. *3nd International Workshop on Digital Mammography*, Elsevier Science, 1996.

[17] Ratan A.L., Maron O., Grimson W.E.L. Lozano-Pérez T.: A Framework for Learning Query Concepts in Image Classification. *Computer Vision and Pattern Recognition Conference*, Fort Collins, CO., June 1999.

[18] Ray S. Page D.: Multiple Instance Regression. Proceedings 18th International Conference on Machine Learning, pp.425-432. San Francisco, CA: Morgan Kaufmann. 2001.

[19] Tsur S.: Data Mining in the Bioinformatics Domain. *VLDB'2000*

[20] Ward N.P., Tomlingson S. Taylor C.J.: Image Analysis of Fundus Photographs: The Detection and Measurements of Exudate Associated with Diabetic Retinopathy. *Ophthal*, Vol. 96, pp.80-86, 1989.

[21] Yang C. Lozano-Pérez T.: Image Database Retrieval with Multiple-Instance Learning Techniques. *Proceeding of 16th ICDE*, San Diego. pp.233-243. 2000.

[22] Zhang Q. Goldman S.A.: EM-DD: An Improved Multiple-Instance Learning Technique. NIPS 2001: pp.1073-1080. 2001.

Yu Xiaoxue Ms. Yu obtained her Bachelor of Science from the Special Class for Gifted Young at the University of Science and Technology of China, majored in Computer Science in 2001. After that she joined the SMA-CS program and obtained Master of Science in 2002. Now she is PhD candidate in this program.

Wynne Hsu Dr. Hsu obtained her Bachelor of Science from the Department of Computer Science at National University of Singapore. After that, she obtained Master of Science from the Department of Computer Science, and PhD from the Department of Electrical and Computer Engineering, both at Purdue University, U.S.A. Dr. Hsu is currently with the Department of Computer Science at the National University of Singapore.

Lee Wee Sun Dr. Lee obtained his Bachelor of Engineering (Hon I) in Computer Systems Engineering from the University of Queensland in 1992. Then he obtained his PhD from the Department of Systems Engineering at the Australian National University in 1996 under Bob Williamson and Peter Bartlett. From 1996 to 1998, Dr. Lee did a post doc under John Arnold and Michael Frater in the School of Electrical Engineering at the Australian Defence Force Academy. Dr. Lee is currently with the Department of Computer Science at the National University of Singapore.

Tomás Lozano-Pérez Tomás Lozano-Pérez is the TIBCO Professor of Computer Science and Engineering at MIT, where he is a member of the Artificial Intelligence Laboratory. Professor Lozano-Pérez has all his degrees (SB '73, SM '76, PhD '80) from MIT in Computer Science. Before joining the MIT faculty in 1981 he was on the research staff at IBM T. J. Watson Research Center during 1977. He has been Associate Director of the Artificial Intelligence Laboratory and Associate Head for Computer Science of MIT's Department of Electrical Engineering and Computer Science. Professor Lozano-Pérez's research has been in robotics (configuration-space approach to motion planning), computer vision (interpretation-tree approach to object recognition), machine learning (multiple-instance learning), medical imaging (computer-assisted surgery) and computational chemistry (drug activity prediction and protein structure determination from NNR X-ray data). He has been co-editor of the International Journal of Robotics Research and a recipient of a Presidential Young Investigator Award from the NSF.