Computerized Medical Imaging and Graphics 42 (2015) 44–50

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Computer-aided diagnosis from weak supervision: A benchmarking study

Melih Kandemir*, Fred A. Hamprecht

Heidelberg University, HCI, Speyerer Str. 6, D-69115 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 4 April 2014
Received in revised form 28 May 2014
Accepted 10 November 2014

Keywords:
Multiple instance learning
Cancer diagnosis
Diabetic retinopathy screening

ABSTRACT

Supervised machine learning is a powerful tool frequently used in computer-aided diagnosis (CAD) applications. The bottleneck of this technique is its demand for fine grained expert annotations, which are tedious for medical image analysis applications. Furthermore, information is typically localized in diagnostic images, which makes representation of an entire image by a single feature set problematic. The multiple instance learning framework serves as a remedy to these two problems by allowing labels to be provided for *groups* of observations, called *bags*, and assuming the group label to be the maximum of the instance labels within the bag. This setup can effectively be applied to CAD by splitting a given diagnostic image into a Cartesian grid, treating each grid element (patch) as an instance by representing it with a feature set, and grouping instances belonging to the same image into a bag. We quantify the power of existing multiple instance learning methods by evaluating their performance on two distinct CAD applications: (i) Barrett's cancer diagnosis and (ii) diabetic retinopathy screening. In the experiments, mi-Graph appears as the best-performing method in bag-level prediction (i.e. diagnosis) for both of these applications that have drastically different visual characteristics. For instance-level prediction (i.e. disease localization), mi-SVM ranks as the most accurate method.

1. Introduction

Advances in image analysis and machine learning gradually make available more robust algorithms for extracting information from data. An appealing application at the intersection of these two disciplines is computer-aided diagnosis which aims to automate disease diagnosis from images [27]. CAD tools have been shown to be useful to aid the pathologist by pointing out important regions in large biopsy tissue images [11], providing decision support by calculating informative metrics such as cell counting [18], and quantifying the disease risk [21].

A major drawback of many CAD algorithms is their demand for fine-grained expert annotations during training. For tumor diagnosis, pathologists need to indicate the tumor regions, and for diabetic retinopathy, small structures such as microaneurysms have to be annonated by opthalmologists. The use of weakly supervised machine learning techniques can drastically reduce the annotation effort, while keeping prediction performance at an acceptable level.

A common characteristic of diagnostic imaging is the locality of discriminative information. For instance, in cancer histology, a small region within a large slide often determines the final grading, and all the remaining slide is redundant. Similarly, in diabetic retinopathy screening, small structures, such as microaneurysms are much richer in diagnostic information than the texture of the entire image. Hence, application of the standard supervised learning setup to these cases would be problematic. Given a diagnostic image, representing it by a single feature vector would require tedious feature engineering, since when standard feature sets are applied, the uninformative areas in the image would overrule the informative ones. On the other hand, dividing the image into small patches, and representing each patch by a feature vector would result in severe class imbalance.

Multiple instance learning (MIL) [19] provides a learning framework that both allows weak supervision and inherently handles the locality of information problem. In MIL, ground-truth labels are available only for groups of observations, called *bags*. A bag with a positive label indicates that there exists at least one observation within that bag, whose label is positive. For a negatively labeled bag, on the other hand, all observations are known to have a negative label. This framework can directly be applied to CAD by defining each diagnostic image (tissue slide or fundus image) as a bag, and each of its regions (e.g. patches in a Cartesian grid) as an instance.

* Corresponding author. Tel.: +49 15430425406.
E-mail addresses: melih.kandemir@iwr.uni-heidelberg.de (M. Kandemir), fred.hamprecht@uni-heidelberg.de (F.A. Hamprecht).

http://dx.doi.org/10.1016/j.compmedimag.2014.11.010
0895-6111/

Diseased cases with local lesions are then represented by a positive bag, and healthy cases by a negative bag.

Even though some previous work reports MIL solutions tailored to specific CAD problems [21,28,29], the utility of a large set of existing MIL approaches in these applications has not yet been evaluated. Furthermore and more importantly, the generalizability of their success on various CAD problems has not yet been quantified. In this paper, we address these two issues by providing a benchmarking study using a large list of MIL methods[1] on two CAD applications that have clearly distinct visual characteristics: (i) diagnosis of Barrett's cancer from H&E stained histology images and (ii) diabetic retinopathy screening from eye fundus images. Among the methods under comparison, mi-Graph [33] outperforms the others in both applications in cancer diagnosis (i.e. prediction of bag labels). On the other hand, in the harder problem of cancer localization (i.e. prediction of instances), mi-SVM [2] gives the highest generalization performance.

2. Prior art

2.1. Cancer diagnosis from histology images

There has been a large volume of studies on application of machine learning methods to histology cancer diagnosis (see [11] for a comprehensive review). Demir et al. [10] propose the classification of brain tumors by constructing graphs from cell topology, and representing the tumor image by a set of graph features. Doyle et al. [8] classify prostate cancer grades from graph-based (e.g. minimum spanning tree of cells), morphological (e.g. nuclear density), and textural features (e.g. Gabor filter responses) using the standard multiclass support vector machine (SVM). Alternatively, Huang et al. [12] show that differential box counting leads to effective prostate cancer grading. Wang [25] demonstrates the successful application of Markov random fields to segmentation of lung tumors. Hang et al. [5] propose a method that combines sparse coding and multiscale histogram intersection kernels for diagnosis of kidney renal carcinoma and glioblastoma. Kandemir et al. [13] perform diagnosis of Barrett's cancer using mi-Graph.

2.2. Automated diabetic retinopathy screening

Agurto et al. [1] introduce an automated diabetic retinopathy screening method that characterizes the texture of regions of interest by their amplitude and frequency properties. Giancardo et al. [9] detect microaneurysms from morphological heuristics, and then apply a standard SVM to predict the disease status. Quellec et al. [22] introduce a content-based image retrieval (CBIR) system for diabetes detection by formulating a probabilistic interpretation of a set of wavelets. In a follow-up study, Quellec et al. [21] improve the state-of-the-art in diabetes detection by extending their CBIR method with multiscale features.

2.3. MIL for computer-aided diagnosis

MIL has comparatively recently started to be used for computer aided diagnosis. Some exemplary studies are as follows. Zhao et al. [32] apply the MILES [6] method to patches of slides of 10 different tissue types. Zhang et al. [30] use GPMIL of Kim et al. [15] for classification of skin biopsies. Xu et al. [28,29] use a multiclass extension of MILBoost [24] for grading of prostate tumors. Quellec et al. [21] build their aforementioned multiscale CBIR method for the MIL setup.

3. The diagnosis pipeline

We use the same automated diagnosis pipeline for both applications. We split a given diagnostic image into a regular grid of patches. We then construct an instance from each patch by extracting a set of features. A group of instances belonging to the same diagnostic image is treated as a bag. The label of the bag is assumed to be +1 if it includes the target disease, and −1 otherwise. Consequently, we predict the disease status of a given image (bag) using one of the MIL methods in comparison. Fig. 1 illustrates the pipeline.

For both applications, we represent an image patch with a set of intensity histogram and texture features as listed in Table 1. For Barrett's cancer diagnosis, we additionally use a set of cell features. We segment cells using supervised pixel classification and watershed transform as described in [13]. We then extract a set of intensity and morphology features from each cell (see Table 2 for the complete list). Finally, we augment the feature vector of each patch by a set of summary statistics of features of cells lying within that patch, as listed in Table 3.

4. Multiple instance learning methods

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ be a data set consisting of N instances, each of which is a D-dimensional feature vector: $\mathbf{x}_i = [x_i^{(1)}, \ldots, x_i^{(D)}]$. The data set is assumed to be partitioned into B bags: $\mathbf{X} = \bigcup_{b=1}^{B} \mathbf{X}_b$, such that $\mathbf{X}_b \bigcap \mathbf{X}_c = \emptyset, \quad \forall b \neq c$, where each bag b consists of N_b instances: $\mathbf{X}_b = [\mathbf{x}_{b1}, \ldots, \mathbf{x}_{bN_b}]$. Let $\mathbf{Y} = [Y_1, \ldots, Y_B]$ be the vector of the corresponding binary bag labels $Y_b \in \{-1, +1\}$. Labels of instances are collected into the vector $\mathbf{y} = [y_1, \ldots, y_N]$, which follows the same partitioning as instances $\mathbf{y} = \bigcup_{b=1}^{B} \mathbf{y}_b$, such that $\mathbf{y}_b \bigcap \mathbf{y}_c = \emptyset, \quad \forall b \neq c$, where $\mathbf{y}_b = [y_{b1}, \ldots, y_{bN_b}]$. Let $\mathcal{B}^+ = \{b | Y_b = +1\}$ and $\mathcal{B}^- = \{b | Y_b = -1\}$ denote sets of positive and negative bags, respectively. Finally, $\mathbb{I}(\cdot)$ denotes the indicator function that gives 1 if its argument is true, and 0 otherwise. The central assumption of the MIL setup is that the label of a bag is the maximum of the labels of the instances in that bag: $Y_b = \max(\mathbf{y}_b)$, which we call as

Table 1
Features extracted from patches of Barrett's cancer histology and fundus images.

Color features	
1	Intensity histogram of RGB channels for 26 bins
Texture features	
2	Mean of local binary pattern histograms of 20 × 20-pixel grids
3	Mean of SIFT descriptors
4	Box count for grid sizes 2, 3, …, 8

Table 2
Features extracted from each segmented cell.

1	Central power sums for exponents 1, 2, 3 and 4
2	Area, radius, perimeter, and roundness of the segment
3	Maximum, mean, and minimum intensity, and intensity, covariance, variance, skewness, and kurtosis within the region and within its 30-pixel-wide belt for each color channel
4	Region axes, principal axes, kurtosis, minimum, maximum, and power sums for exponents 1, 2, 3, 4

Table 3
Features extracted from cells located within each Barrett's cancer image patch.

Minimum, maximum, mean, standard deviation, skewness, and kurtosis of features (given in Table 2) of all healthy and cancer cells in a patch

[1] The source code of the MIL methods in our comparison list is available under: http://hci.iwr.uni-heidelberg.de/Staff/mkandemi/MILBundle.tar.gz.