Computational and Structural Biotechnology Journal xxx (2018) xxx–xxx

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Machine Learning Methods for Histopathological Image Analysis

Daisuke Komura *, Shumpei Ishikawa

Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

A R T I C L E I N F O

Article history:
Received 30 August 2017
Received in revised form 3 December 2017
Accepted 14 January 2018
Available online xxxx

Keywords:
Histopathology
Deep learning
Machine learning
Whole slide images
Computer assisted diagnosis
Digital image analysis

A B S T R A C T

Abundant accumulation of digital histopathological images has led to the increased demand for their analysis, such as computer-aided diagnosis using machine learning techniques. However, digital pathological images and related tasks have some issues to be considered. In this mini-review, we introduce the application of digital pathological image analysis using machine learning algorithms, address some problems specific to such analysis, and propose possible solutions.

1. Introduction

Pathology diagnosis has been performed by a human pathologist observing the stained specimen on the slide glass using a microscope. In recent years, attempts have been made to capture the entire slide with a scanner and save it as a digital image (whole slide image, WSI) [1]. As a large number of WSI are being accumulated, attempts have been made to analyze WSIs using digital image analysis based on machine learning algorithms to assist tasks including diagnosis.

Digital pathological image analysis often uses general image recognition technology (e.g. facial recognition) as a basis. However, since digital pathological images and tasks have some unique characteristics, special processing techniques are often required. In this review, we describe the application of digital pathological image analysis using machine learning algorithms, and its problems specific to digital pathological image analysis and the possible solutions. Several reviews that have been published recently discuss histopathological image analysis including its history and details of general machine learning algorithms [2–7]; in this review, we provide more pathology-oriented point of view.

Since the overwhelming victory of the team using deep learning at ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012 [8], most of the image recognition techniques have been replaced by deep learning. This is also true for pathological image analysis [9–11]. Therefore, even though many techniques introduced in this review are related to deep learning, most of them are also applicable for other machine learning algorithms.

2. Machine Learning Methods

Fig. 1 shows typical steps for histopathological image analysis using machine learning. Prior to applying machine learning algorithms, some pre-processing should be performed. For example, when cancer regions are detected in WSI, local mini patches around 256 × 256 are sampled from large WSI. Then feature extraction and classification between cancer and non-cancer are performed in each local patch. The goal of feature extraction is to extract useful information for machine learning tasks. Various local features such as gray level co-occurrence Matrix (GLCM) and local binary pattern (LBP) have been used for histopathological image analysis, but deep learning algorithms such as convolutional neural network [9,10,12–14] starts the analysis from feature extraction. Features and classifiers are simultaneously optimized in deep learning and features learned in deep learning often outperforms other traditional features in histopathological image analysis.

Machine learning techniques often used in digital pathology image analysis are divided into supervised learning and unsupervised learning. The goal of supervised learning is to infer a function that can map the input images to their appropriate labels (e.g. cancer) well using training data. Labels are associated with a WSI or an object in WSIs. The algorithms for supervised learning include support vector machines, random forest and convolutional neural networks. On the other hand, the goal of unsupervised learning is to infer a function that can describe hidden structures from unlabeled images. The tasks include clustering, anomaly detection and dimensionality reduction. The algorithms for unsupervised learning include k-means, autoencoders and principal component analysis. There are derivatives from these two learning such as semi-supervised learning and multiple instance learning, which are described in Section 4.2.2.

* Corresponding author.
E-mail address: dkom.gpat@mri.tmd.ac.jp (D. Komura).

https://doi.org/10.1016/j.csbj.2018.01.001

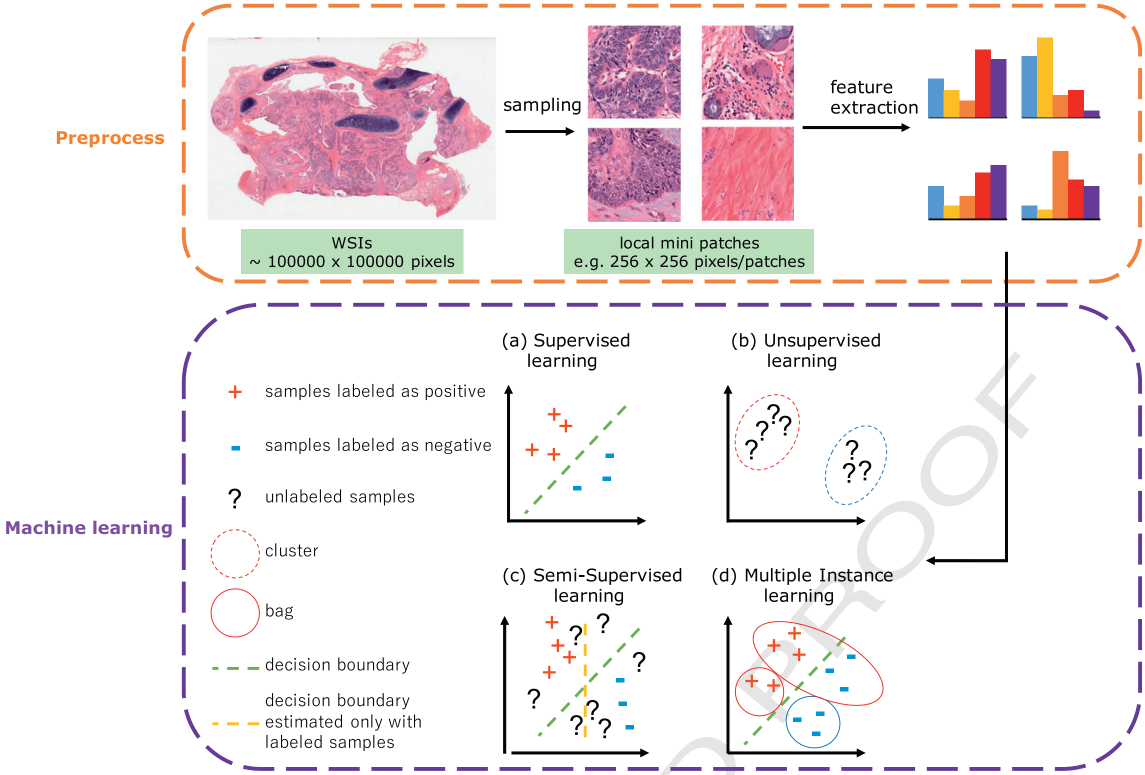

Fig. 1. Typical steps for machine learning in digital pathological image analysis. After preprocessing whole slide images, various types of machine learning algorithms could be applied including (a) supervised learning (see Section 2), (b) unsupervised learning (see Section 2), (c) semi-supervised learning (see Section 4.2.2), and (d) multiple instance learning (see Section 4.2.2). The histopathological images are adopted from The Cancer Genome Atlas (TCGA)[34].

3. Machine Learning Application in Digital Pathology

3.1. Computer-assisted Diagnosis

The most actively researched task in digital pathological image analysis is computer-assisted diagnosis (CAD), which is the basic task of the pathologist. Diagnostic process contains the task to map a WSI or multiple WSIs to one of the disease categories, meaning that it is essentially a supervised learning task. Since the errors made by a machine learning system reportedly differ from those made by a human pathologist [15], classification accuracy could be improved using CAD system. CAD may also lead to the reduce variability in interpretations and prevent overlooking by investigating all pixels within WSIs.

Other diagnosis-related tasks include detection or segmentation of Region of Interest (ROI) such as tumor region in WSI [16,17], scoring of immunostaining [11,18], cancer staging [15,19], mitosis detection [20,21], gland segmentation [22–24], and detection and quantification of vascular invasion [25].

3.2. Content Based Image Retrieval

Content Based Image Retrieval (CBIR) retrieves similar images to a query image. In digital pathology, CBIR systems are useful in many situations, particularly in diagnosis, education, and research [26–31]. For example, CBIR systems can be used for educational purposes by students and beginner pathologists to retrieve relevant cases or histopathological images of tissues. In addition, such systems are also helpful to professional pathologists, particularly when diagnosing of rare cases.

Since CBIR does not necessarily require label information, unsupervised learning can be used [30]. When label information is available, supervised learning approaches could learn better similarity measure than unsupervised learning approaches [28,29] since the similarity between histopathological images may differ by definition. However, preparing sufficient number of labeled data can be a serious problem as will be described later.

In CBIR, not only accuracy but also high-speed search of similar images from numerous images are required. Therefore, various techniques for dimensionality reduction of image features such as principal component analysis and compact bilinear pooling [32], and fast approximate nearest neighbor search such as kd-tree and hashing [33] are utilized for high speed search.

3.3. Discovering New Clinicopathological Relationships

Historically, many important discoveries concerning diseases such as tumor and infectious diseases have been made by pathologists and researchers who have carefully and closely observed pathological specimens. For example, *H. pylori* was discovered by a pathologist who was examining the gastric mucosa of patients with gastritis [34]. Attempts have also been made to correlate the morphological features of cancers with their clinical behavior. For example, tumor grading is important in planning treatment and determining a patient's prognosis for certain types of cancer, such as soft tissue sarcoma, primary brain tumors, and breast and prostate cancer.

Meanwhile, thanks to the progress in digitization of medical information and advance in genome analysis technology in recent years, large amount of digital information such as genome information, digital pathological images, MRI and CT images has become available [35]. By analyzing the relationship between these data, new clinicopathological relationships, for example, the relationship between the morphological characteristic and the somatic mutation of the cancer, can be found [35,36]. However, since the amount of data is enormous, it is not realistic for pathologists and researchers to analyze all the relationships manually by looking at the specimens. This is where the machine learning technology comes in. For example, Beck et al. extracted texture information from pathological images of breast cancer and analyzed with L1 -

Please cite this article as: Komura D, Ishikawa S, Machine Learning Methods for Histopathological Image Analysis, Comput Struct Biotechnol J (2018), https://doi.org/10.1016/j.csbj.2018.01.001