Pattern Recognition Letters 42 (2014) 115–121

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Cancer diagnosis by nuclear morphometry using spatial information ☆,☆☆

Hu Huang [a], Akif Burak Tosun [a,*], Jia Guo [a], Cheng Chen [a], Wei Wang [a], John A. Ozolek [b], Gustavo K. Rohde [a,c,d]

[a] Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[b] Department of Pathology, Children's Hospital of Pittsburgh, Pittsburgh, PA 15224, USA
[c] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[d] Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 29 July 2013
Available online 20 February 2014

Keywords:
Set classification
Naïve Bayes
Majority voting
Thyroid lesion classification
Cancer diagnosis

ABSTRACT

Methods for extracting quantitative information regarding nuclear morphology from histopathology images have been long used to aid pathologists in determining the degree of differentiation in numerous malignancies. Most methods currently in use, however, employ the *naïve Bayes* approach to classify a set of nuclear measurements extracted from one patient. Hence, the statistical dependency between the samples (nuclear measurements) is often not directly taken into account. Here we describe a method that makes use of statistical dependency between samples in thyroid tissue to improve patient classification accuracies with respect to standard *naïve Bayes* approaches. We report results in two sample diagnostic challenges.

1. Introduction

Given the prominent role of nuclear structure changes in cancer cells [1–3], numerous researchers have made use of quantitative nuclear structure measurements to describe automated methods for classifying different lesions. Automated systems aimed at detection and diagnosis (grading) of cancerous tissues from histopathology images have been described for diagnosing breast cancer [4–8], thyroid cancer [9–11], prostate cancer [12,13], liver cancer [14] and colon cancer [15], to name a few. In these methods the following general strategy is typically used (see Fig. 1). First, images of tissue specimens, usually obtained via surgical procedures and stained with a particular stain (e.g. hematoxylin and eosin), are taken using transmission light microscopy, for example. After appropriate preprocessing (e.g. color unmixing), the nuclei are segmented and numerical features describing their morphological characteristics (e.g. size, perimeter, texture features) are extracted and used to train a classifier which is capable of determining whether a set of nuclei extracted from a particular individual can be classified as benign or malignant, or given a differential diagnosis.

One prominent characteristic of many of the methods that use nuclear morphometry to grade different kinds of cancers is that classification is performed using the *naïve Bayes* method whereby each nuclear structure (represented by a set of numerical features) is often classified independently from one another [16,17,11]. The set of nuclei extracted from a patient is then usually classified by using the majority voting (MV), or taking the most common class assignment, or perhaps by using different moments (e.g. mean, variance) of the distribution of nuclei. Thus any statistical dependency, such as correlation for example, between nearby structures is discarded. Several attempts to capture the spatial information between nearby cells from microscopic images have been made by using the graph theory [18,13]. In these works the x, y position of each nuclear structure in a field of view is used to generate a neighborhood graph which, together with average nuclear features, is used in an attempt to differentiate different classes. Information regarding the intricate distribution of the numerical features describing each structure, as well as co-dependencies between these in nearby nuclei, however, are often not used explicitly.

Our goal in this methodological note is to demonstrate that any amount of statistical dependency between the morphological characteristics of nearby nuclei can be utilized to improve the classification accuracy of methods usually employed for cancer diagnosis and differentiation. It is well known that cells in living tissues utilize several mechanisms (e.g. autocrine or paracrine) to 'communicate' with one another. Given that well established cell communication

☆ This work was supported in part by NIH awards R21GM088816, R01GM090033.
☆☆ This paper has been recommended for acceptance by C. Luengo.
* Corresponding author. Address: 5000 Forbes Avenue, Hamerschlag Hall C-119, Pittsburgh, PA 15213, USA. Tel.: +1 412 268 8379.
E-mail addresses: huh@andrew.cmu.edu (H. Huang), tosun@cmu.edu (A.B. Tosun), ozolja@upmc.edu (J.A. Ozolek), gustavor@cmu.edu (G.K. Rohde).

http://dx.doi.org/10.1016/j.patrec.2014.02.008
0167-8655/

Fig. 1. A typical flowchart of histopathology image-based computer-aided diagnosis.

mechanisms exist, it could then be possible that the morphological information of a given nucleus could depend (statistically speaking) on the morphology of nearby nuclei. Here we present evidence that indeed numerical features of nuclei are more correlated to features extracted from nearby nuclei rather than those of distant nuclei, and that this difference is statistically significant. We then describe a method that utilizes any dependency present to augment the accuracy of classification (e.g. benign vs. malignant) in comparison with the *naïve Bayes* strategy (e.g. majority voting).

We note that the idea of classifying sets of samples (nuclei), rather than individual samples, is not new and has been studied in pattern recognition domains recently. In multiple instance learning (MIL) algorithms, for example, [19,20], the learner receives a set of bags (each containing more than one sample) that are labeled positive or negative. Here each bag is labeled, and not each sample. In MIL algorithms, however, a bag is labeled negative if all the instances in it are negative, but a bag is labeled positive if there is at least one instance in it which is positive. Other than MIL algorithms, [21], for example, investigated different instance learning methods, focusing on the classifier model construction. Under the same context, [17] proposed a *K*-nearest neighbor method for group-based classification by combining a MV scheme and a pooling scheme. They indicate that knowing a set of test samples that belong to the same, but unknown, class can be used to effectively reduce the individual Bayes error rate. Similar approaches that combine individual classification methods with the MV strategy were also investigated in the high-throughput applications [22,23] and revealed an improved classification performance compared to those not using MV strategy. In a similar manner, the method we describe below makes use of the spatial x, y position of nuclei in a field of view to exploit their dependency for augmented classification accuracies. We demonstrate the performance of our approach by classifying three types of thyroid lesions from 78 patients.

The remainder of this paper is structured as follows. In Section 2, we describe the mathematical model for the set classification problem, and show the relationship between the MV strategy and the likelihood ratio test (LRT) strategy. We then describe a method that is able to utilize 'sets of nuclei' extracted from image neighborhoods instead of individual nuclei. We note the new method does not require a specific ordering within each sub-group. Section 3 describes the computational procedures we utilized to demonstrate the application of our approach. Section 4 presents experimental results comparing the several computational strategies involved. Finally, summary and conclusions are offered in the last section of this document.

2. Bayesian framework

Let x_j^i be a d-dimensional numerical feature vector describing the jth nucleus of the ith patient, and let $X_i = \{x_1^i, x_2^i, \ldots, x_{n_i}^i\}$ describe the set of feature vectors pertaining to all nuclei belonging to the ith patient. Given a set of nuclear measurements X_i, the objective in pathology problems is to determine the class label $y \in \{y_1, y_2, \ldots, y_k\}$ (for a problem with k gradings or classes) for this set of measurements. The *maximum a posterior* (MAP) criterion can be used to estimate the label of the set X_i via:

$$y_{X_i} = \arg\max_y p(y|X_i) = \arg\max_y \frac{p(X_i|y)p(y)}{p(X_i)} = \arg\max_y \frac{p(x_1^i, x_2^i, \ldots, x_{n_i}^i|y)p(y)}{p(x_1^i, x_2^i, \ldots, x_{n_i}^i)}. \quad (1)$$

For a two-class problem, the label could be simply determined by comparing the posterior probabilities, given by

$$\frac{p(y=y_1|X_i)}{p(y=y_2|X_i)} = \frac{p(X_i|y=y_1)p(y_1)}{p(X_i|y=y_2)p(y_2)} \quad (2)$$

and testing whether this ratio is smaller or greater than one. By assuming the prior probability of each class is equal, i.e. $p(y_1) = p(y_2)$ (when no *a priori* information regarding incidence is available), the likelihood ration test (LRT) [24] can be further simplified as

$$\mathcal{L} = \log\left(\frac{p(X_i|y=y_1)}{p(X_i|y=y_2)}\right) = \log\left(p(X_i|y=y_1)\right) - \log\left(p(X_i|y=y_2)\right) \quad (3)$$

$$y_{X_i} = \begin{cases} y_1, & \mathcal{L} > 0 \\ y_2, & \mathcal{L} < 0. \end{cases} \quad (4)$$

Computing the joint conditional probability $p(X_i|y) = p(x_1^i, x_2^i, \ldots, x_{n_i}^i|y)$ is often difficult given the low number of samples in comparison with the number of dimensions ($d \times n_i$) that this would involve. The *naïve Bayes* assumption is then often used to overcome this problem. In this approach, it is assumed that the samples (nuclei) are independent from one another, i.e. $p(x_1^i, x_2^i, \ldots, x_{n_i}^i|y) = \prod_{j=1}^{n_i} p(x_j^i|y)$. Under this assumption, the log-likelihood ratio in Eq. (3) can be computed as

$$\begin{aligned}\mathcal{L} &= \log\left(p(X_i|y=y_1)\right) - \log\left(p(X_i|y=y_2)\right) \\ &= \log\left(\prod_{j=1}^{n_i} p\left(x_j^i|y=y_1\right)\right) - \log\left(\prod_{j=1}^{n_i} p\left(x_j^i|y=y_2\right)\right) \\ &= \sum_{j=1}^{n_i} \log\left(p\left(x_j^i|y=y_1\right)\right) - \sum_{j=1}^{n_i} \log\left(p\left(x_j^i|y=y_2\right)\right). \end{aligned} \quad (5)$$

Another approach that is often used in these situations is the MV strategy [25]. The main idea is to classify each sample in the case individually by using a chosen classifier, label each sample accordingly, and then assign the label with the majority of votes as the final label for the case (patient). In order to analyze the connection between MV and LRT, let the output of an individual classifier be $C(x_j^i) \in [-\infty, \infty]$, and define an indicator function as

$$I(p) = \begin{cases} 1, & \text{if } p \geqslant 0 \\ 0, & \text{if } p < 0, \end{cases} \quad (6)$$

where 1 denotes class y_1, and 0 refers to class y_2. Then the class label for the case X_i is determined by calculating the numbers of samples belonging to each class

$$y_{X_i} = \begin{cases} y_1, & \text{if } \sum_{j=1}^{n_i} I(C(x_j^i)) \geqslant \frac{n_i}{2} \\ y_2, & \text{if } \sum_{j=1}^{n_i} I(C(x_j^i)) < \frac{n_i}{2}. \end{cases} \quad (7)$$

Note that if $C(x_j^i)$ is defined as the log-ratio of the posterior probabilities, then we could obtain similar functions as in Eqs. (3) and (4),

$$\begin{aligned}\mathcal{MV} &= \sum_{j=1}^{n_i} I\left[\log\left(\frac{p\left(x_j^i|y=y_1\right)}{p\left(x_j^i|y=y_2\right)}\right)\right] \\ &= \sum_{j=1}^{n_i} I\left[\log\left(p\left(x_j^i|y=y_1\right)\right) - \log\left(p\left(x_j^i|y=y_2\right)\right)\right] \end{aligned} \quad (8)$$

$$y_{X_i} = \begin{cases} y_1, & \text{if } \mathcal{MV} \geqslant \frac{n_i}{2} \\ y_2, & \text{if } \mathcal{MV} < \frac{n_i}{2}. \end{cases} \quad (9)$$