Artificial Intelligence in Medicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading

Shahnorbanun Sahran[a], Dheeb Albashish[b,*], Azizi Abdullah[a], Nordashima Abd Shukor[c], Suria Hayati Md Pauzi[c]

[a] Pattern Recognition Research Group, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Malaysia
[b] Computer Science Department, Prince Abdullah Bin Ghazi Faculty of Information Technology, Al-Balqa Applied University, Jordan
[c] Department of Pathology, University Kebangsaan Malaysia Medical Center, 56000 Batu 9 Cheras, Malaysia

ARTICLE INFO

Article history:
Received 29 April 2017
Received in revised form 2 April 2018
Accepted 7 April 2018

Keywords:
Prostate histopathological image
Tissue components
Ensemble classification
Feature selection
SVM-RFE
Absolute cosine
Redundancy

ABSTRACT

Objective: Feature selection (FS) methods are widely used in grading and diagnosing prostate histopathological images. In this context, FS is based on the texture features obtained from the lumen, nuclei, cytoplasm and stroma, all of which are important tissue components. However, it is difficult to represent the high-dimensional textures of these tissue components. To solve this problem, we propose a new FS method that enables the selection of features with minimal redundancy in the tissue components.
Methodology: We categorise tissue images based on the texture of individual tissue components via the construction of a single classifier and also construct an ensemble learning model by merging the values obtained by each classifier. Another issue that arises is overfitting due to the high-dimensional texture of individual tissue components. We propose a new FS method, SVM-RFE(AC), that integrates a Support Vector Machine-Recursive Feature Elimination (SVM-RFE) embedded procedure with an absolute cosine (AC) filter method to prevent redundancy in the selected features of the SV-RFE and an unoptimised classifier in the AC.
Results: We conducted experiments on H&E histopathological prostate and colon cancer images with respect to three prostate classifications, namely benign vs. grade 3, benign vs. grade 4 and grade 3 vs. grade 4. The colon benchmark dataset requires a distinction between grades 1 and 2, which are the most difficult cases to distinguish in the colon domain. The results obtained by both the single and ensemble classification models (which uses the product rule as its merging method) confirm that the proposed SVM-RFE(AC) is superior to the other SVM and SVM-RFE-based methods.
Conclusion: We developed an FS method based on SVM-RFE and AC and successfully showed that its use enabled the identification of the most crucial texture feature of each tissue component. Thus, it makes possible the distinction between multiple Gleason grades (e.g. grade 3 vs. grade 4) and its performance is far superior to other reported FS methods.

1. Introduction

The common method used to confirm the diagnosis of prostate cancer is a detailed examination and grading of tissue slides. These slides are usually obtained by needle biopsy of the prostate, which is then imaged with a microscope using post-haematoxylin and eosin (H&E) staining. Tumours are graded based on the Gleason grading procedure [1], which consists of five grades from least aggressive (level 1) to most aggressive (level 5). These grades are primarily determined by careful analysis of the texture of the prostate tissue [2]. However, the grading process has limitations due to the very similar features of the glandular tissue components between grades 3 and 4 [1] (e.g. lumen, epithelial cell nuclei, epithelial cell cytoplasm and stroma). Pathologists find the task of manually diagnosing prostate cancer to be cumbersome, due to the intricate nature of detecting anomalies by observing prostate tissue under a microscope [3]. Another commonly reported problem is that diagnoses are highly dependent on the skill of the pathologist.

* Corresponding author.
E-mail addresses: shahnorbanun@ukm.edu.my (S. Sahran), bashish@bau.edu.jo (D. Albashish), azizia@ukm.edu.my (A. Abdullah), nordashima@ppukm.ukm.edu.my (N.A. Shukor), su_hayati@ppukm.ukm.edu.my (S. Hayati Md Pauzi).

https://doi.org/10.1016/j.artmed.2018.04.002
0933-3657/

The above problems can be mitigated by the use of machine learning approaches in a computer-aided diagnosis (CAD) system for grading prostate cancer. To do so, first, a set of texture features is obtained from regional tissue images or tissue components [4]. However, it is well known that not every feature is used for grading. Also, the utilisation of the multiple texture feature yields a high-dimensional feature vector size, which leads to a crucial issue [5]. In the context of high-dimensionality, machine learning algorithms, such as the support vector machine (SVM), experience the curse of dimensionality [6], which culminates in overfitting [5]. These issues can be addressed using feature selection (FS) methods, which are excellent for determining which features are characterised by minimum redundancy [7].

FS is classified as a filter, wrapper, or embedded approach [7], depending on how it is combined with the learning model. SVM recursive feature elimination (SVM-RFE) [8] is an embedded method introduced by Guyon for gene selection and cancer classification. This method is more robust in overfitting data than the wrapper and filter methods [8]. Although SVM-RFE has many advantages, its main disadvantage is the redundancy of its selected features [9]. A few approaches have been reported by the authors in [9,10]. Yoon et al. [9] first use SVM-RFE to rank the relevancy of the features, then rank the feature-based redundancy criterion. However, a disadvantage of this approach is that the nonlinear combination of relevance and redundancy criteria can result in fewer redundant features with minimal relevancy, which decreases performance. To address this shortcoming, a new FS method was proposed that integrates the SVM-RFE [8] embedded method with the absolute cosine filter method [11] to realise high classification performance with minimum redundancy among the selected features.

The remainder of this paper is organised as follows: in Section 2, we introduce the background of prostate-cancer grading CADs and review the FS literature. We describe algorithms used in the ensemble framework, feature extraction and feature selection and classification in Section 3 and in Section 4, we present our experimental results and discussion. We draw our conclusions in Section 5.

2. Related work

Below, we discuss previous work in digital pathology, including computer-aided prognosis for prostate cancer and a comparison of FS methods in a hybridised SVM-RFE with other filter methods.

2.1. Review of existing prostate CAD

The Gleason patterns in histopathological images differ in their texture and the configuration of their tissue components. When devising prostate CADs to discriminate between various Gleason grades (patterns) in tissue images, the main question to consider is – what are the best features that can be used to differentiate between Gleason patterns? Based on their feature extraction procedures, published studies addressing prostate CADs can be categorised into one of two approaches. The first is texture-based CAD, which characterises Gleason patterns based on the spatial variations in the pixel values. The Gleason patterns can be distinguished based on texture characteristics, such as smooth, coarse and fine [4]. Second are tissue-structure-based CAD systems, which utilise morphological features obtained from a specific tissue component in the region of interest (ROI), examples being the nuclei and lumen.

The majority of CADs that rely on texture features utilise the Haralick texture features [12] obtained from a co-occurrence matrix to train single or ensemble machine learning classifiers to classify test images [13–15]. For instance, DiFranco et al. [13] presented an ensemble framework to increase the ability to distinguish between normal and abnormal prostate cancer patterns. In this study, the authors independently extracted Haralick texture features from each colour channel (RGB and CIEL*a*b*). Utilising random forest FS and linear SVM, the features showed an AUC of 94.8% when classifying tiles into normal and abnormal categories. Despite the fact that the tiles were automatically chosen, those consisting of more than a single pattern were eliminated from the training dataset. Tabesh et al. [15] classified tissue images based on their Haralick texture features, colour channel variance and nuclei arrangement into low-level and high-level Gleason grades, with 81.0% and 97% accuracies for a normal vs. cancerous classifications, respectively. However, these highly accurate results were obtained based only on small spots on a tissue microarray [16]. Diamond et al. [14] characterised tissue patch texture based on the Haralick texture [12,17] and morphometric features, assuming that abnormal tissues exhibit smaller areas of associated lumen. The authors reported an accuracy level of 79.3% when assessing the algorithm within the sub-regions of eight tissue images (using 40× magnification). However, the patch was quite small, at 100 × 100, and the installed validation schemes were not properly detailed. The texture features utilised in previously reported texture-based CAD systems are generic and computed from image pixels. Moreover, no domain knowledge is used in the prostate histopathological images.

Tissue-structure-based CAD systems (as opposed to texture-based CAD) have also been explored by scholars. Nguyen et al. [3] and Naik et al. [18] showed that morphological features, such as the shape of the lumen and nuclei tissue components, can be used to discriminate between Gleason patterns. In his preliminary work, Nguyen et al. [3] utilised low-level domain information to classify the lumen and nuclei tissue components, followed by obtaining multi-shaped features associated with glands and lumen. From a dataset consisting of 26 images, the authors reported an accuracy rate of 85.5% when classifying grades 3 and 4. However, the authors also reported a few limitations, the main one being the occlusion of lumen tissue components in certain tissue images [4]. Naik et al. [18] utilised a Bayesian classifier to classify pixels according to their respective colours into basic tissue components. Then, they extracted the shape of the lumen and glandular shape features. From 44 images, the reported diagnosis and grading results for Grade 3 vs. grade 4 was 95%, for grade 3 vs. benign was 86% and for grade 4 vs. benign was 92%. Texture features obtained from the distribution of tissue components have also been used [19]. For instance, Khurd et al. [20] used the texture features of the nuclei tissue component and reported an accuracy of 91.5% when categorising 75 images into grades 3 vs. 4. One of the main limitations of their study was the need to manually annotate the nuclei objects.

The shortcoming of state-of-the-art tissue structure-based CAD systems is that they are reliant on the presence of tissue components, although some basic tissue components, such as lumens, are occluded by cytoplasm [3]. Therefore, accurate high-level features measurement cannot always be achieved. We proposed an ensemble framework in a previous work [21] based on the texture features of the tissue components (mainly, lumen, cytoplasm, nuclei and stroma). In that paper, we reported the utilisation of a set of 41 Gleason grade 3s and 56 Gleason grade 4s and a resulting AUC performance of 93.59% for grade 3 vs. grade 4, which is regarded as the most complex classification task. In this work, we used the ensemble framework to determine the impact of the proposed FS in PCa grading.

2.2. Feature selection methods

Most Gleason grading CAD systems are defined by a high-dimensional feature space that is highly related to the number of samples, which is referred to as the overfitting (curse-of-

Please cite this article in press as: Sahran S, et al. Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.002