Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A hierarchical classifier based on human blood plasma fluorescence for non-invasive colorectal cancer screening

Felipe Soares [a,*], Karin Becker [a], Michel J. Anzanello [b]

[a] Institute of Informatics – Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, Porto Alegre, RS, Brazil
[b] Department of Industrial Engineering – Universidade Federal do Rio Grande do Sul, Av. Osvaldo Aranha, 99–5° andar, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 7 June 2017
Received in revised form 15 August 2017
Accepted 12 September 2017

Keywords:
Colorectal cancer
Hierarchical classification
Fluorescence spectroscopy
Support vector machine
Human blood plasma
SVM-RFE

ABSTRACT

Colorectal cancer (CRC) a leading cause of death by cancer, and screening programs for its early identification are at the heart of the increasing survival rates. To motivate population participation, non-invasive, accurate, scalable and cost-effective diagnosis methods are required. Blood fluorescence spectroscopy provides rich information that can be used for cancer identification. The main challenges in analyzing blood fluorescence data for CRC classification are related to its high dimensionality and inherent variability, especially when analyzing a small number of samples. In this paper, we present a hierarchical classification method based on plasma fluorescence to identify not only CRC, but also adenomas and other non-malignant colorectal findings that may require further medical investigation. A feature selection algorithm is proposed to deal with the high dimensionality and select discriminant fluorescence wavelengths. These are used to train a binary support vector machine (SVM) in the first level to identify the CRC samples. The remaining samples are then presented to a one-class SVM trained on healthy subjects to detect deviant samples, and thus non-malignant findings. This hierarchical design, together with the one class-SVM, aims to reduce the effects of small samples and high variability. Using a dataset analyzed in previous studies comprised of 12,341 wavelengths, we achieved much superior results. Sensitivity and specificity are 0.87 and 0.95 for CRC detection, and 0.60 and 0.79 for non-malignant findings, respectively. Compared to related work, the proposed method presented a better accuracy, required fewer features, and provides a unified approach that expands CRC detection to non-malignant findings.

1. Introduction

Colorectal cancer (CRC) is still one of the major causes of death by cancer [1], but significant advances on mortality reduction have been achieved by modern CRC screening methods [2]. Indeed, early CRC detection drastically improves the 5 year survival rate [3], while detection and removal of other non-malignant conditions, such as adenomas, can even prevent further development to CRC. Adenomas are polyps (i.e. tissue growth) that can progress to CRC, reason why they are usually removed when detected [4]. The development of cost-effective screening methods able to accurately identify not only colorectal cancer samples, but also other non-malignant related conditions, is of utmost interest for effective CRC screening programs, improving the outcomes of preventive and curative medical interventions.

Colonoscopy is the gold standard method for CRC detection, with reported sensitivity ranging between 92% and 99% [5]. Despite its effectiveness, its coverage is affected by high costs, low capacity of medical centers, as well as psychological factors related to its invasiveness nature [6–9], which hinder a wider participation of the target population. Concerning non-invasive methods, stool-based tests such as Guaiac fecal occult blood test (gFOBT) and fecal immunochemical test (FIT) are among the most common CRC screening techniques [2,10], but their predictive power for positive CRC is limited. Thus, investigation of new non-invasive diagnostic methods is encouraged [8,11–14], especially blood-based tests, since patients are more keen to accept them, based on a preference of not handling stool samples [6].

More recently, fluorescence spectroscopy has been employed for cancer identification in tissues [15] and body fluids [16,17]. Fluorescence spectroscopy is a well-known tool for chemical analysis, which provides detailed molecular information [18,19] at relative low cost [20], thus suitable for non-invasive CRC screening. One

* Corresponding author.
E-mail addresses: felipe.soares@inf.ufrgs.br (F. Soares), karin.becker@inf.ufrgs.br (K. Becker), anzanello@producao.ufrgs.br (M.J. Anzanello).

http://dx.doi.org/10.1016/j.artmed.2017.09.004
0933-3657/

of the key issues in fluorescence analysis is the definition of excitation and emission wavelengths, which are the features derived from fluorescence spectroscopy. Therefore, among the main challenges in analyzing fluorescence data from biological systems are the high dimensionality inherent to fluorescence spectroscopy, and the usual small number of available samples and their high variability. High dimensionality is intrinsically characterized by the existence of many noisy features that can lead to overfitting [21]. Hence, it requires methods for the identification of proper features for classification, in order to obtain quality classification results and develop a scalable solution. The latter may affect multiclass classification performance, because classifiers must deal with more complex patterns to distinguish the samples [22].

Lualdi et al. [23] developed a feasibility study of blood plasma fluorescence for CRC detection. They selected an excitation wavelength of 405 nm and emission recorded in the range of 430–700 nm, achieving sensitivity and specificity results of 80% and 50%, respectively. Lawaetz et al. [3] significantly improved these results by experimenting with various excitation and emission wavelengths. However, fluorescence data acquisition in a range of excitation and emission wavelengths is time consuming, and may inflate the dataset with irrelevant information for the classification task. Therefore, a dimension reduction pre-processing step is required to produce sound classifiers [24]. The usual methods for dimension reduction are feature extraction and feature selection. In the former, the original features are combined together to generate a new reduced set of features that maximizes a given criterion (e.g. explained variance). In the latter, a reduced subset of the original features is selected, discarding the irrelevant or redundant ones for the classification task. As a result, only the most discriminant features are used to build the classification model.

Regarding feature selection, Lualdi et al. [23] applied *t*-test to discard the non-significant emission wavelengths to build univariate classifiers based on fluorescence intensity. The reported results demonstrate the potential of fluorescence spectroscopy for cancer identification, and how results rely on the choice of proper features. In regards to feature extraction for CRC identification, related work [3,25] employed Parallel Factor Analysis (PARAFAC) for extracting interpretable features to train PLS-DA (Partial Least Squares Discriminant Analysis) binary classifiers. Despite little performance improvement was achieved, they have shown that spectroscopy-derived features provide enough information to build classifiers for different target classes (e.g. CRC, adenomas, healthy). However, they did not integrate these independent binary classifiers. Thus, there is an opportunity to develop a unifying and encompassing solution that identifies all these CRC-related conditions, where feature selection has the main purpose of improving classification results.

In this paper, we propose a classification framework based on fluorescence spectroscopy from human blood plasma to identify CRC samples and other non-malignant findings. The striking characteristics of our solution are: a) the inclusion of a feature selection pre-processing step that can improve overall accuracy and reduce sample acquisition time; and b) the adoption of a hierarchical classifier to identify non-malignant findings, which decomposes the classification problem according its intrinsic subclass structure, and thus leads to more homogeneous classes. The proposed unifying framework can be the basis of a quicker, cost-effective and scalable non-invasive CRC screening method. We use publicly available blood plasma fluorescence datasets to build and test the classifiers, such that our method can be compared to previous studies [3,25].

With regard to prior attempts of CRC and non-malignant findings classification, a major difference is that we address this problem as a multi-class classification task. Existing solutions are limited to individual binary classifiers (e.g. CRC versus healthy subjects, or CRC versus non-CRC). Multi-class problems are usually more challenging than binary ones, since defining precise class boundaries may be more difficult. Additionally, the small number of samples, high dimensionality and high variability further accentuate this challenge. We propose the use of a two level hierarchical design to solve this multi-class classification problem. As depicted in Fig. 1, it is composed of a binary SVM at the first level, and of a one-class SVM in the second level. The first level binary SVM classifier has the role of separating the well-characterized CRC samples from the rest, whereas the second one aims at handling non-malignant findings samples as outliers with regard to healthy patients.

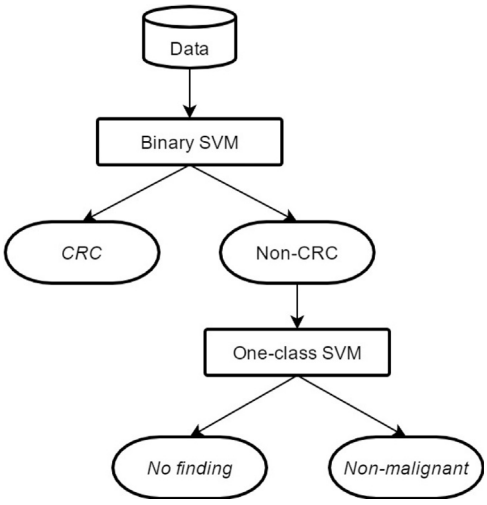

Fig. 1. Hierarchical classifier.

Another novel contribution of our paper is the process employed for dimension reduction. Unlike previous works [3,23,25], we adopt an iterative multivariate feature selection process. A Support Vector Machine – Recursive Feature Elimination (SVM-RFE) algorithm [21] is used to rank and select the appropriate features, leading to a simplified, yet more accurate model, with a reduced number of features.

The proposed approach can provide invaluable information for individuals that may develop colorectal cancer in the future. We foresee our method as a preliminary step that helps selecting patients who need colonoscopy or other invasive procedures, avoiding unnecessary exams. Our method can increase participation rate and coverage in CRC screening programs, due to its non-invasive nature and relative low cost.

The remaining of this paper is structured as follows. Section 2 presents information about fluorescence spectroscopy, the dataset used in this study, the SVM algorithm, and the detailed steps to build the hierarchical classifier. Results of the proposed approach are presented in Section 3, and compared to other classification strategies and previous works on the same dataset. Conclusions are drawn in Section 4, together with future work.

2. Materials and methods

In this section, we briefly present the fluorescence spectroscopy technique, the dataset used in this study, the classification algorithm and the proposed method for hierarchical classification.

2.1. Fluorescence spectroscopy

Fluorescence is one of the categories of luminescence (i.e. emission of light) that occurs when an excited molecule called fluorophore emits energy in form of light to return to its ground state [26]. Fluorescence can be interpreted as a three-stage process. First, a compound is illuminated at an excitation wavelength, absorbing

Please cite this article in press as: Soares F, et al. A hierarchical classifier based on human blood plasma fluorescence for non-invasive colorectal cancer screening. Artif Intell Med (2017), http://dx.doi.org/10.1016/j.artmed.2017.09.004