Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Predicting the risk of *squamous dysplasia and esophageal squamous cell carcinoma* using minimum classification error method

Motahareh Moghtadaei [a,1], Mohammad Reza Hashemi Golpayegani [a,*], Farshad Almasganj [b,2], Arash Etemadi [c,d,3], Mohammad R. Akbari [e,f,3], Reza Malekzadeh [d]

[a] *Complex Systems and Cybernetic Control Lab., Faculty of Biomedical Engineering, Amirkabir University of Technology (Tehran Polytechnic), P.O. Box 1591634311, Tehran, Iran*
[b] *Speech Lab., Faculty of Biomedical Engineering, Amirkabir University of Technology (Tehran Polytechnic), P.O. Box 1591634311, Tehran, Iran*
[c] *Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA*
[d] *Digestive Disease Research Center, Shariati Hospital, Tehran University of Medical Sciences, P.O. Box 1411713135, Tehran, Iran*
[e] *Womens College Research Institute, Womens College Hospital, University of Toronto, Toronto, Canada*
[f] *Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*

ARTICLE INFO

Article history:
Received 3 June 2013
Accepted 18 November 2013

Keywords:
Early detection of cancer
Risk factors
Cancer prediction
Squamous dysplasia
Esophageal squamous cell carcinoma
Classification

ABSTRACT

Early detection of squamous dysplasia and esophageal squamous cell carcinoma is of great importance. Adopting computer aided algorithms in predicting cancer risk using its risk factors can serve in limiting the clinical screenings to people with higher risks. In the present study, we show that the application of an advanced classification method, the Minimum Classification Error, could considerably enhance the classification performance in comparison to the logistic regression model and the variable structure fuzzy neural network, as the latest successful methods. The results yield the accuracy of 89.65% for esophageal squamous cell carcinoma, and 88.42% for squamous dysplasia risk prediction.

1. Introduction

Esophageal cancer (EC) is the eighth common cancer globally and is the sixth most common cause of cancer death worldwide. Two most common types of EC are squamous cell carcinoma (ESCC) and adenocarcinoma (EAC). EAC is the most common type in western countries while ESCC is still the predominant type worldwide, especially in the developing countries [14].

ESCC is the squamous cell carcinoma mostly seen in the middle or upper one-third of the esophagus [14]. In recent years, incidence rates for ESCC have been steadily declining in several western countries but it is increasing in certain Asian areas that stretch from northern Iran through the central Asian republics to north-central China [14].

The five-year relative survival of ESCC is less than 20% in the United States [15]. In Golestan Province, five-year relative survival of ESCC is only 3.3% with the median survival of only 7 months [7]. For these reasons, diagnosis of ESCC and squamous dysplasia, in very early stages in order to prevent tumor formation, local invasion and metastasis is of great importance [26].

Although squamous dysplasia can be detected by endoscopy and biopsy [4], but endoscopic screening of the whole population is not practical due to huge costs needed and time required for screening (endoscopy and biopsy) of the whole population. Any new method that can distinguish the people with higher risks of squamous dysplasia, and ESCC, is valuable because it can decrease the need for clinical screenings. The complexity of the tumor system concerning different levels of gene, molecular, cellular, tissue, organ, body and population interacting via various complicated signal transduction pathways [16] makes it difficult to detect people with higher cancer risks.

Mathematical modeling is a very useful tool for estimating the complicated, unknown dependency between risk factors as inputs and the possibility of squamous dysplasia, and ESCC initiation as the output.

In a previous study by Etemadi et al. in 2012 [7], a logistic regression model has been exploited for modeling and prediction of squamous dysplasia, and ESCC using their risk factors.

The classification methods that try to calculate the probability distributions, like the method used by Etemadi et al. to

[*] Corresponding author. Tel./fax: +98 21 64542389.
E-mail address: mrhashemigolpayegani@aut.ac.ir (M.R. Hashemi Golpayegani).
[1] Tel./fax: +98 21 64542389.
[2] Tel.: +98 21 64542372; fax: +98 21 66495655.
[3] Tel.: +98 21 82415300; fax: +98 21 82415400.

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.11.011

estimate the regression model often, do not lead to an optimal performance especially in biological applications. That is because in complex biological applications the *a priori* probabilities that express the related uncertainties before the data is taken into account, and the state conditional densities of classes are unknown. On the other hand, usually the data cannot represent the posterior probability distributions either. In such models, the estimated probabilities are an approximation of the true probabilities, and the occurred modeling errors prevent maximum A-posteriori probability (MAP) rule to be implemented as accurate as it could potentially operate [16].

To avoid this problem, we recently suggested a non-statistical variable structure fuzzy neural network (FNN) as an approximator which does not need to primarily estimate probabilities [21]. To optimize the model, we adopt a hybrid global chaotic optimization algorithm (COA). This method is proved to have a higher performance than the logistic regression model, in the ESCC and dysplasia risk prediction [21].

Moreover, in this paper, a different strategy is adopted, and we are going to face the mentioned shortness of statistical classifiers via a compensated approach so-called the minimum classification error (MCE) classifier. This is an efficient approach to estimate the Gaussian mixture model (GMM) parameters to effectively be employed in sorting the subjects with higher risks of squamous dysplasia, and ESCC. However, we will see that this method is more efficient in comparison to the mentioned hybrid COA. The continuous search region of the COA algorithm finally led to the optimum response but is very time consuming. So, the computational time is reduced via the suggested hybrid method [21]. The MCE method does not face such problem and is computationally more efficient.

In the following section, the database used in this study is first described; in Section 3, the MCE algorithm for classification is briefly described. Subsequently, the classification results obtained via MCE are reported in Section 4; finally, the discussion and conclusion are included in Sections 5 and 6, respectively.

2. Data description

Two datasets are used in this study. The first dataset is relating to ESCC, and includes 300 biopsy-proven ESCC cases and 571 age and sex-matched neighborhood controls. This dataset is from the Golestan Case-Control Study that was conducted from 2003 to 2007 [22].

The second dataset is relating to dysplasia and includes 26 individuals with dyplastic lesions and 698 controls. This dataset is collected from individuals visiting Atrak Clinic, a gastroenterology research clinic in Gonbad City, Golestan Province, between 2002 and 2007. Video endoscopy with Lugol's iodine staining, questionnaire and biopsies helped to develop this dataset [22].

The conduct of studies performed to obtain the dataset including risk factors were reviewed and approved by the Institutional Review Boards of Tehran University Digestive Disease Research Center (DDRC), the US National Cancer Institute (NCI) and the International Agency for Research on Cancer (IARC) [22].

In each of the cases, the data set includes all known risk factors associated to dysplasia and ESCC known in the Golestan Province [1,2,12,13,22].

The risk factors are described below:

(i) *Age*

(ii) *Place of residence*: Place of residence of subjects contributed in this study include the counties of Gonbad, Minoodasht, Kalaleh, Azadshahr, and Ramian in eastern Golestan province [22].

(iii) *Ethnicity*: Approximately half of the residents of the study area are of Turkmen ethnicity, and the rest Persians, Kurds, Turks, and others [22].

(iv) *Tobacco smoking*: Cumulative use (average intensity multiplied by duration of use) of tobacco is considered [22].

(v) *Opium use*: Cumulative use (average intensity multiplied by duration of use) of opium is considered [22].

(vi) *Socio-economic status*: A socio-economic score is assigned to each subject considering education level, and relatives and family structure. A wealth score is also assigned to each subject considering house ownership, house structure, house size, number of people living together in the current house, ownership of household appliances, and the duration of owning these appliances, and the most recent occupation of subjects, using a multiple correspondence analysis (MCA) [13].

(vii) *Oral health*: Frequency of brushing teeth is used as the most important oral hygiene factor [1].

(viii) *Family history*: The data relating the family history contains information on all of the first- and second-degree relatives and first cousins including the vital status of these family members and all occurrences of esophageal cancer and other cancers. In addition, current age, age at diagnosis of cancer, site of cancer, age of death, clinical and pathological diagnosis of cancer were recorded for all first-degree relatives. The presence of parental consanguinity was also recorded for cases and controls [2].

(ix) *Tea temperature*: The tea temperature degree, estimated by the time interval between tea being poured and drunk, was also recorded for each case [12].

(x) *Water source*: Subject's water source, access to piped water, and years having access to piped water were recorded [22].

In general, in order to express the mentioned risk factors in terms of numbers, there is a total of 21 numbers recorded for each subject in the ESCC and dysplasia datasets.

To better describe the datasets, two important histograms for ESCC and dysplasia datasets are demonstrated in Figs. 1 and 2, respectively. Presenting the histograms of all of the data is not feasible due to the large number recorded data.

3. The MCE method

The MCE algorithm is a type of discriminant training algorithms. In the commonly used model-based classification approaches, in which individual statistical models are constructed for different classes of data, the model parameters are found in a way to minimize the error between the model and the real probability distribution. So, the classification decision rule in such algorithms does not appear in the overall training phase of the models, and it does not necessarily lead to minimizing the overall classification error. In the MCE training algorithm, in addition to the modeling error, the final classification result also roles in the final optimization of the models' parameters [16,27]. Consider an input vector x_n, and its target class y_n, $\{x_n, y_n\}_{n=1}^{N}$. In the general form of the MCE training algorithm, in the case of K classes, the classifier makes its decision by the following decision rule

$$y = k \text{ if } g_{y=k}(x_n, \Lambda) = \max_{\text{for all } i \epsilon K} g_{y=i}(x_n, \Lambda) \tag{1}$$

where y is the decision of the classifier for input x_n, or equivalently

$$y = k \text{ if } g_{y=k}(x_n, \Lambda) - \max_{\text{for all } i \neq k} g_{y=i}(x_n, \Lambda) > 0 \tag{2}$$

where $g_{y=i}(x_n, \Lambda)$ is the discriminant function by which the score of the assignment of x_n to class i is evaluated, and Λ is the