RESEARCH COMMUNICATION

Feature Selection Methods for Optimizing Clinicopathologic Input Variables in Oral Cancer Prognosis

Siow-Wee Chang1*, Sameem Abdul Kareem1, Thomas George Kallarakkal2, Amir Feisal Merican Aljunid Merican3, Mannil Thomas Abraham4, Rosnah Binti Zain2

Abstract

The incidence of oral cancer is high for those of Indian ethnic origin in Malaysia. Various clinical and pathological data are usually used in oral cancer prognosis. However, due to time, cost and tissue limitations, the number of prognosis variables need to be reduced. In this research, we demonstrated the use of feature selection methods to select a subset of variables that is highly predictive of oral cancer prognosis. The objective is to reduce the number of input variables, thus to identify the key clinicopathologic (input) variables of oral cancer prognosis based on the data collected in the Malaysian scenario. Two feature selection methods, genetic algorithm (wrapper approach) and Pearson’s correlation coefficient (filter approach) were implemented and compared with single-input models and a full-input model. The results showed that the reduced models with feature selection method are able to produce more accurate prognosis results than the full-input model and single-input model, with the Pearson’s correlation coefficient achieving the most promising results.

Keywords: Feature selection - oral cancer prognosis - genetic algorithm - Pearson’s correlation coefficient

Asian Pacific J Cancer Prev, 12, 2659-2664

Introduction

The mortality rate for oral cancer is high (at approximately 50%) because the cancer is always discovered late in its development. The well known risks associated with this cancer include smoking, alcohol consumption, tobacco use, and betel quid chewing. Besides risks factors, there are other factors associated with oral cancer such as viral infection, genetic factors, diet, and poor oral hygiene (Jeffries & Foulkes, 2001; Reichart, 2001; Sunnitha & Gabriel, 2004; Mehrotra & Yadav, 2006). The World Health Organization (WHO) expects a worldwide rise in oral cancer incidence in the next few decades due to high smoking prevalence and increasing cases of unhealthy diet. Almost two-thirds of oral cancer occurs in developing countries for example India, South East Asia, and Brazil, and this geographic variation probably reflects the prevalence of specific environmental influences and risk habits (Oliveira et al., 2008).

There are various clinical and pathological data which are used by the clinicians for oral cancer prognosis. Clinical data refers to the signs and symptoms directly observable by the clinicians, the examples are, size of primary lesion, site of lesion, clinical neck node, clinical staging, metastasis, and so on. While, pathological data relates to the results obtained from the laboratory examination and the parameters are pathological staging, number of neck nodes, invasions, tumor size and thickness. In this research, both clinical and pathological data are used, and they are referred to as clinicopathologic data.

The common problem that is associated with medical dataset is small sample size with large variable sets. It is time-consuming and costly to obtain large amount of samples in medical research and the samples are usually inconsistent, incomplete or noisy in nature. Furthermore, if the sample size is small and the numbers of variables are big, it will cause over-fitting problems. Over-fitting occurs when there are too many parameters relative to the number of samples. High accuracy and reliable estimation is needed in medical diagnosis and prognosis where the subsequent decisions have serious consequences on patients. Thus, a simple predictive model with reduced variables is more efficient as compared to a full-model prediction.

In this research, a wrapper feature selection method, genetic algorithm, has been selected and the results are compared and validated with a filter method - Pearson’s correlation coefficient. Next, the features (clinicopathologic variables) selected from both methods are tested and compared with the single-input and full-input models using the logistic regression for the