RESEARCH COMMUNICATION

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

Siow-Wee Chang[1*], Sameem Abdul Kareem[1], Thomas George Kallarakkal[2], Amir Feisal Merican Aljunid Merican[3], Mannil Thomas Abraham[4], Rosnah Binti Zain[2]

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 (Jefferies & Foulkes, 2001; Reichart, 2001; Sunnitha & Gabrial, 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

*[1]Department of Artifical Intelligence, Faculty of Computer Science and Information Technology & Bioinformatics Division, [3]Bioinformatics Division, Institute of Biological Science, Faculty of Science, [2]Department of Oral Pathology and Oral Medicine and Periodontology, Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya, [5]Hospital Tunku Ampuan Rahimah, Ministry of Health Malaysia, Kuala Lumpur, Malaysia *For correspondence: changsiowwee@yahoo.com*

classification of 3-year oral cancer survival.The proposed methods are tested with oral cancer prognosis dataset collected locally here in Malaysia. The objectives are to produce a simpler predictive model with reduced variables, and hence to overcome the over-fitting problems in small medical dataset.

Oral cancer starts in the mouth, also called the oral cavity. The oral cavity includes the lips, the inside lining of the lips and cheeks (buccal mucosa), the teeth, the gums, the front two-thirds of the tongue, the floor of the mouth below the tongue, the bony roof of the mouth (hard palate), and the area behind the wisdom teeth (retromolar trigone) (Omar et al., 2006).

In Malaysia, Indians are more susceptible to oral cancer and Indian women face the greatest risk, this might be related to their oral habits of betel quid chewing. According to Figure 1, adapted from the Malaysian Cancer Statistics, Peninsular Malaysia 2006 (American Cancer Society, 2010), tongue cancer is listed as the sixth top most cancer (4.6%) in Indian male, and mouth cancer is listed as the fourth top most cancer (7.3%) in Indian female. Although oral cancer is not listed as the top ten most occurring cancer in Malaysia, the high mortality rate related to this cancer has resulted in the need to improve its survival rate.

In Malaysia, oral cancer incidence rate is the highest (71.6%) for individual above 50 years old. From the gender perspective, both genders carry almost equal percentage of cancer incidence, with male at 45.9%, and female at 54.1%. Tongue cancer has the highest incidence rate when compared to cancers in other parts of the mouth.

A multicentre study from Malaysia (Mustafa et al., 2007) indicates that among 156 oral cancer patients, the risk habits that was most commonly practiced was betel quid chewing (59.9%), followed by smoking (36.1%) and alcohol consumption (35.2%). Apart from socio-demographic risk factors such as age, ethnic, gender, habits, the other factors associated with oral cancer are virus infection, gene factors, diet, and oral hygiene. Another study (Tan et al., 2005) indicates that in comparing oral cancer patients against healthy, non-cancer patients, it was found that frequent intake of vegetables were higher among those who did not have cancer (83%) as opposed to those who have (70%).

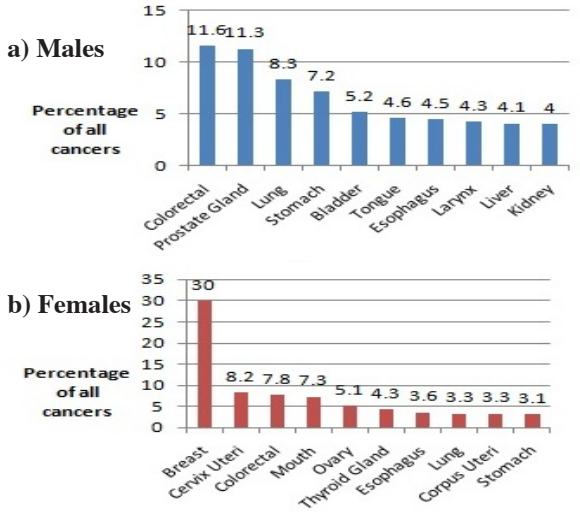

Figure 1. Ten Most Frequent Cancers in Indians, Peninsular Malaysia 2006. *Modified from Malaysia Cancer Statistics, Peninsular Malaysia 2006 (Omar et al., 2006)

The oral cancer prognosis (OCP) dataset used in this research is obtained from the Malaysian Oral Cancer Database and Tissue Bank System (MOCDTBS) coordinated by the Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya (Zain et al., 2009). The dataset consists of oral cancer cases collected from the participating hospitals from all over Malaysia. From this database, 31 samples were selected based on the completeness of the clinicopathologic data. The data consists of information for socio-demographic data for example age, gender, and habits; clinicopathologic data such as primary sites, clinical and pathological Tumor-Node-Metastasis (TNM) stage, nodal status, tumor size, invasion status, types of treatments, survival information and others.

Due to the vast numbers of clinicopathologic variables and the small sample size, it is important to implement feature selection methods in the proposed model to avoid over-fitting. The aim is to minimize the number of clinicopathologic inputs and thus to reduce the time and costs needed for oral cancer prognosis. In this research, feature selection methods will be implemented into the OCP dataset to choose the most optimal clinicopathologic variables. Next, the selected clinicopathologic variables will be used for the prediction of 3-year oral cancer survival, that is, either the patient survives or dies 3 years after diagnosis.

Materials and Methods

Feature selection is used to select the inputs which are most significant in the modeling process, in order to produce more accurate outputs. The purpose of feature selection is to reduce the number of inputs in the modeling process, but retain the accuracy of the outputs as compared to the full-input model. Thus, this can produce a more predictive and cost effective model. This is important especially in medical research where fewer inputs means lower test and diagnosis/prognosis costs.

Feature selection can be classified into three main groups, which are filter, wrapper and embedded methods. Filter methods rank the variables by some chosen criterion, and select the variables with highest criteria. This method, however, is independent of any algorithm. Whereas, the wrapper methods evaluate the variables in subsets and use the heuristic search methods for an optimal subset. The embedded method is built into a classifier to search for a subset and it is specific to the learning algorithm (Talavera, 2005; Saeys et al., 2007).

There are various feature selection techniques that have been implemented, for example, in (Song et al., 2005), a couple of feature selection methods i.e. genetic algorithm, decision tree and correlation coefficient computation are proposed with ANFIS and Adaboost in order to reduce the computational overhead and enhance the system performance. Their results showed that ANFIS with the feature selection system performed better than ANFIS

full-input system with ANFIS-decision tree achieving the highest positive predicted value (98%).

Principal component analysis (PCA) has been proposed as a feature selection tool for clinical pattern recognition analysis for thyroid cancer and cervical cancer (Zhang, 2007). The PCA was applied on the multiple layer perceptron artificial neural networks (MLP ANN). The researchers proved that the accuracy rate of the MLP ANN based on the PCA input selector was improved as compared to leave-one-out cross-validation method. They claimed that they achieved 100% classification rate with the proposed method.

Sun et al. (2007) proposed a new feature selection algorithm named as I-RELIEF. I-RELIEF combines the advantages of both filter and wrapper methods. It approximates the leave-one-out accuracy of a nearest-neighbour classifier, thus, it addresses the issues of feature correlation and removal of redundant features. It is used to identify a hybrid signature through the combination of both genetic and clinical markers. The results showed that the hybrid signature model outperformed other models for breast cancer prognosis.

Genetic Algorithm

Genetic algorithms (GA) were formally introduced in the United States in the 1970s by John Holland at the University of Michigan. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms (also known as evolutionary computation) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (Cordon et al., 2001).

The algorithm starts with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness, those with best fitness are selected through the process of crossover (exchanging properties) and mutation (changes in the properties). This is repeated until some conditions (for example number of populations or improvement of the best solution) is satisfied (Obitko, 1998).

The GA is proposed as a feature selection method the small sample size of medical data in this oral cancer prediction research. In this case, the solutions of the GA will form the clinicopathologic variables that will subsequently be used in the oral cancer prognosis and the output will indicate how well the solutions can predict oral cancer survival. The GA was run and tested using Matlab's Genetic Algorithm and Direct Search toolbox, version 7.

Solution Encoding

In the feature subset selection problem, a solution is specific feature subset that can be encoded as a string of n binary digits (bits). Each feature is represented by binary digits of 1 or 0. If a bit is equal to 1, the feature is selected; consequently, if a bit is equal to 0, the feature is not selected. For example, in the oral cancer prognosis dataset, if the solution is 0110010000100000 string of 16 binary digits, it indicates that features 2, 3, 6, and 11 are selected as the feature subset.

Initial population

The initial population is generated randomly to select a subset of variables (solutions). If the variables are all different, the subset is included in the initial population. If not, it generates again until an initial population with desired size has been created.

Fitness function

The function is use to classify between two groups, which are alive and death. The error rate of the classification will be calculated using a 10-fold cross-validation. The fitness function is the final error rate obtained. The subset of variables with the lowest error rate will be selected.

Selection

Selection is a process to select the parent chromosome from the population to reproduce the next generation. In this study, the roulette wheel selection (Obitko, 1998) is chosen where the fittest individuals have a higher chance of being selected than weaker ones.

Crossover

The crossover function that used in this study is crossover scattered. It creates a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent, and combines the genes to form the child. The crossover fraction is set at 0.5, it means that 50% of the children other than elite individuals are crossover children. In addition, the crossover function is set to ensure that it do not return the repeated variables.

Mutation

Same as crossover, the mutation is set to ensure that it do not return the repeated variables. Mutation uniform is used, where the algorithm selects a fraction of an individual for mutation, where each entry has a probability rate of being mutated. Next, the algorithm replaces each selected entry by a random number selected uniformly from the range for that entry. The mutation rate is set at 0.3.

Stopping criteria

The stopping criteria used in this study is the number of generations and time limit. The number of generation is set at 100 and the time limit is 600s, whichever occur first for the GA to stop.

Performance measures

In a medical prognosis problem, a person with positive condition (alive) who is predicted as alive is termed a true positive (TP), whereas a person with positive condition (alive) who is predicted as negative is termed a false negative (FN). On the other hand, a person with negative condition (dead) who is predicted as positive is termed as false positive (FP), while a person with negative condition (dead) who is predicted as negative is termed as true

Table 1. Original Clinicopathologic Input Features

No.	Name	Description
1	Age	Age at diagnosis
2	Eth	Ethnicity
3	Gen	Gender
4	Smoke	Smoking habit
5	Drink	Alcohol drinking habit
6	Chew	Quid chewing habit
7	Site	Primary site of tumor
8	Subtype	Subtype and differentiation for SCC*
9	Inv	Invasion front
10	Node	Neck nodes
11	PT	Pathological tumor staging
12	PN	Pathological lymph nodes
13	PM	Pathological metastasis
14	Stage	Overall stage
15	Size	Size of tumor
16	Treat	Type of treatment

*SCC, Squamous cell carcinoma

Table 2. Accuracy, Sensitivity, Specificity and AUC for Single-Input Models

Features	Accuracy	Sensitivity	Specificity	AUC
Age	64.5	70.0	54.6	0.71
Eth	67.7	95.0	18.2	0.61
Gen	64.5	100.0	0.0	0.58
Smoke	64.5	100.0	0.0	0.52
Drink	64.5	85.0	27.3	0.56
Chew	64.5	85.0	27.3	0.56
Site	71.0	90.0	36.4	0.66
Subtype	64.5	100.0	0.0	0.51
Inv	64.5	100.0	0.0	0.63
Node	64.5	100.0	0.0	0.57
PT	71.0	90.0	36.4	0.73
PN	67.7	80.0	45.5	0.65
PM	64.5	100.0	0.0	0.54
Stage	64.5	90.0	18.2	0.61
Size	64.5	5.0	27.3	0.62
Treat	64.5	80.0	36.4	0.62

negative (TN). Figure 2 listed the confusion matrix for oral cancer prognosis.

Several measures were used to evaluate and validate the performance of the proposed model. The measures are accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curve. The true performance of the model is defined as the area under the ROC curve (AUC). Sensitivity is the true positive conditions divided by all the living patients. The specificity is the true negative conditions divided by all the dead patients. Accuracy is the proportion of true results in the samples, the higher the accuracy, the better the model is. The ROC curve is a plot of sensitivity versus (1 - specificity) for different test results. The area calculated under the ROC curve is termed as area under curve (AUC).

		Actual conditions	
		Alive (Positive)	Dead (Negative)
Predicted outcomes	Alive (Positive)	True positive (TP)	False positive (FP)
	Dead (Negative)	False negative (FN)	True negative (TN)

Figure2. Confusion Matrix for Oral Cancer Prognosis

Pearson's Correlation Coefficient

Pearson's correlation coefficient, r, is used to see if the values of two variables are associated. It measures the strength and the direction of a linear relationship between two variables. It was developed by Karl Pearson and is sometimes referred to as Pearson's product moment correlation coefficient. The mathematical formula for computing r between two variables of x and y, with n sample size, is denoted as (MathBits.com, 2010):

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{nS_x S_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

The correlation coefficient is a number between -1 and 1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. A positive correlation indicates a direct relationship, and a negative correlation indicates an inverse relationship between two variables. If there is no relationship between the predicted values and the actual values, the correlation coefficient is 0 or very low. Thus, the higher the correlation coefficient, the better the input variable is.

Results

We have implemented the proposed feature selection methods for our oral cancer prognosis dataset. The aim is to build a simpler and more accurate predictive model for a 3-year prognosis for oral cancer patients. The original dataset consists of 16 clinicopathologic input features and 31 samples. The original input features are shown as in Table 1.

For the GA feature selection, we tested the model with the different combinations of number of input ranged from 3-input to 7-input. The purpose was to find out the most optimum set of features that can best predict the survival of oral cancer. It was observed that models of 4-input, 5-input, 6-input and 7-input had the same error rate while the 3-input had the higher error rate. Hence, the 4-input model was selected, since it is the simplest model with lower error rate.

Next, the Pearson's correlation coefficient for every input variable and 3-year prognosis is calculated. The features that have the highest r and lowest p-value are selected. In this case, we chose top 4 inputs only in order to tally with the comparison with the GA method, which are, namely: age, drink, invasion and PN.

Next, we tested the selected features with logistic regression for classification and compared the ROC curves (Figure 3) for both of the feature selection models (reduced models) with single-input models, and the full-input model. The predictive performance for each model is measured in terms of accuracy, sensitivity, specificity and area under the ROC curve (AUC) as listed in Table 2 and Figure 3.

From the results obtained in Table 2 and Figure 3, we can see that all reduced models outperformed the single-input models and the full-input model. The reduced

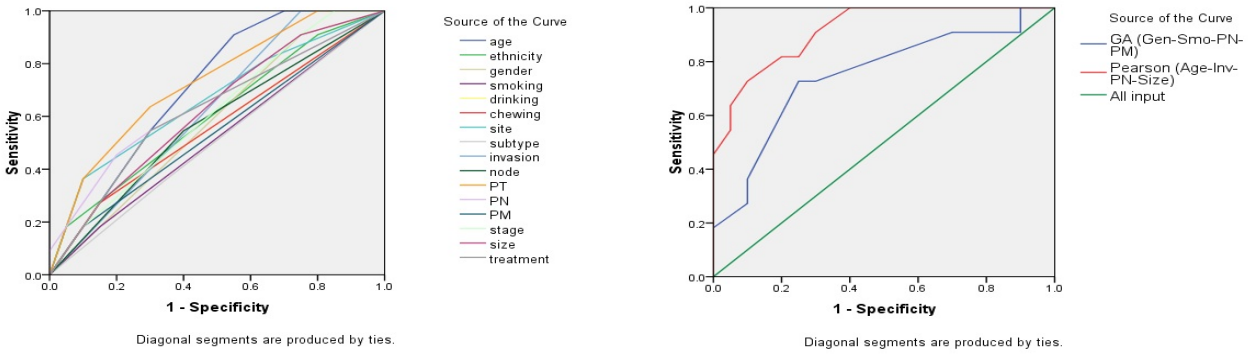

Figure 3. ROC Curves. a) Single-input Models; b) Reduced and Full-Input Models

model with Pearson's correlation coefficient achieved the highest accuracy of 83.9%, sensitivity of 95%, specificity of 63.6%, and AUC of 0.92. The AUC for the full-input model is 0.50 and the best AUC for the single-input model is 0.73 (PT). The poor results generated from the full-input model might be due to the over-fitting problems. The promising results have shown that the models with feature selection methods are able to produce more accurate results as compared to those without, with the Pearson's correlation coefficient outperformed the GA method.

The inputs selected by the Pearson's method are age, invasion, PN and size. Our findings are in accordance with some previous studies which have proved that these inputs are important prognosis factor for oral cancer survival. In (Chen et al., 2007), they had proved that prognosis was the worst in elderly subject in Taiwan. The same goes for (Oliveira et al., 2008) and (Razak et al., 2010) studies, they found that patients aged over 60 years had poorer prognosis if compared to younger patients. In (Walker et al., 2003), depth of invasion is one of the most important predictors of lymph node metastasis in tongue cancer. The TNM staging is use world widely as a prognosis factor for cancer, in which lymph node metastasis (PN) is a significant prognostic factor for oral cancer (Hiratsuka et al., 1997; Li et al., 2005).Walker et al. (2003) and Capilla et al. (2007) showed that size of tumor is one of the factors most associated with oral cancer mortality. However, more testing and verifications need to be done in order to find out the most promising prognostic factors for oral cancer.

Discussion

In this research, we compared two types of feature selection methods, which are genetic algorithm and Pearson's correlation coefficient. From the classification results obtained from the oral cancer prognosis model, we found that the reduced models with feature selection method performed better than full-model and single-input model, with the Pearson's method outperformed the rests.

The four features selected by the Pearson's correlation coefficient selection methods are age, invasion, PN and size. These results are in agreement with some previous studies but there are still other prognostic factors which have been investigated and proved by the others, namely, betel quid chewing, drinking, histopathological status, primary sites of tumor and biomarkers (Reichart, 2001; Sunnitha & Gabriel 2004; Song et al., 2005; Cheng et al., 2007; Saeys et al., 2007; Zhang, 2007; Oliveira et al., 2008). Thus, the results obtained in this research still require further study in order to find out the most relevant and accurate prognostic factors for oral cancer. Our future works include obtaining more oral cancer samples locally and include the biomarkers in our study.

The sample size for oral cancer prognosis data is very small, thus, the feature selection method is a must to reduce the number of clinicopathologic input variables to avoid the over-fitting problem. feature selection methods are suitable for medical research which has the key features of limited time, cost and tissue samples.

Acknowledgements

This study is supported by the University of Malaya Research Grant (UMRG) with the project number RG026-09ICT. The authors would like to thank staff of Oral & Maxillofacial Surgery Department, Oral Pathology Diagnostic Laboratory staff, OCRCC staff, Faculty of Dentistry, and staff of ENT department, Faculty of Medicine, University of Malaya for the preparation of the dataset and related documents for this project.

References

American Cancer Society (2010). Oral Cavity and Oropharyngeal Cancer. [Online]. Available from: http://www.cancer.org/Cancer/OralCavityandOropharyngealCancer/DetailedGuide/oral-cavity-and-oropharyngeal-cancer-what-is-oral-cavity-cancer [Accessed 20th October 2010].

Capilla MV, Olid MNR, Gaya MVO, Botella CR, Ruiz VB (2007). Factors related to survival from oral cancer in an Andalusian population sample (Spain). *Med Oral Patol Oral*, **12**, 518-23.

Chen, PH, Shieh TY, Ho PS, et al (2007). Prognostic factors associated with the survival of oral and pharyngeal carcinoma in Taiwan. *BMC Cancer*, **7**, 101.

Cordon O, Herrera F, Hoffmann F, Magdalena L (2001). Genetic Fuzzy Systems-Evolutionary Tuning and Learning of Fuzzy Knowledge Bases, World Scientific Publishing.

Hiratsuka H, Miyakawa A, Nakamori K, et al (1997). Multivariate analysis of occult lymph node metastasis as a prognostic indicator for patients with squamous cell carcinoma of the oral cavity. *Cancer,* **80**, 351–6.

Jefferies S, Foulkes WD (2001). Genetic mechanisms in squamous cell carcinoma of the head and neck. *Oral Oncol*, **37**, 115-26.

Li XM, Di B, Shang YD, Li J, Cheng JM (2005). Clinicopathologic

features and prognostic factors of cervical lymph node metastasis in oral squamous sell sarcinoma. *Chinese J Cancer*, **24**, 208-1.

MathBits.com (2010). Correlation Coefficient. [Online]. Available from: http://mathbits.com/mathbits/tisection/statistics2/correlation.htm [Accessed 10th November 2010].

Mehrotra R, Yadav S (2006). Oral squamous cell carcinoma: etiology, pathogenesis and prognostic value of genomic alterations. *Indian J Cancer*, **43**, 60-6.

Mustafa WMW, Ghani WMN, Karen NLP, et al. (2007). Survival of Oral cancer Patients in Malaysia- A Multi-Centre Audit. *J Dent Res*, **86** (Special Issue B), 101(SEA).

Obitko, M. (1998). Introduction to Genetic Algorithm. [Online]. Available from: http://www.obitko.com/tutorials/genetic-algorithms/ga-basic-description.php [Accessed 31st May 2010].

Oliveira LR, Ribeiro-Silve A, Costa, JPO, et al. (2008). Prognostic factors and survival analysis in a sample of oral squamous cell carcinoma patients. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*, **106**, 685-95.

Omar ZA, Ali ZM, Tamin NSI (2006). Malaysian Cancer Statistics - Data and Figure, Peninsular Malaysia 2006, National Cancer Registry, Ministry of Health Malaysia.

Razak AA, Saddki, N, Naing NN, Abdullah N (2010). Oral cancer survival among Malay patients in Hospital Universiti Sains Malaysia, Kelantan. *Asian Pac J Cancer Prev*, 11, 187-91.

Reichart PA (2001). Identification of risk groups for oral precancer and cancer and preventive measures. *Clin Oral Invest*, **5**, 207-13.

Saeys Y, Inza I, Larranaga P (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507-17.

Song HJ, Lee SG, Park GT (2005). A Methodology of Computer Aided Diagnostic System on Breast Cancer. IEEE Conference on Control Applications, Torronto, Canada.

Sunitha C, Gabriel R (2004). Oral Cancer At A Glance. *The Internet J Dental Sci*, 1(2).

Sun Y, Goodison S, Li J, Liu L, Farmerie W (2007). Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, **23**, 30-7.

Talavera L (2005). An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering. 6th International Symposium on Intelligent Data Analysis, Madrid, Spain.

Tan MN, Ghazali SHM, Tan CE, Ghani WMN, Zain RB (2005). Habitual and nutritional factors for oral cancer among Malaysians. *J Dent Res,* **84** (B), 028 (SEA).

Walker DM, Boey G, McDonald LA (2003). The pathology of oral cancer. *Pathology*, **35**, 376–83.

Zain RB, Ghani WMN, Razak IA, et al. (2009). Building partnership in oral cancer research in a developing country - processes and barriers. *Asian Pac J Cancer Prev*, **10**, 513-8.

Zhang YX (2007). Artificial neural networks based on principal component analysis input selection for clinical p attern recognition analysis. *Talanta*, **73**, 68-75.