Asian Journal of Information Technology 17 (2): 160-166, 2018
ISSN: 1682-3915

Multi-Target Regression Prediction on Cervical Cancer for Evaluation of Predictive Performance Measures

[1]S.G. Fashoto, [1]A.S. Metfula, [1]B.B. Matsebula and [2]B.Y. Fashoto
[1]Department of Computer Science, University of Swaziland, Kwaluseni, Swaziland
[2]Baylor College of Medicine Children's Foundation, Kampala, Uganda

Abstract: Multi-target prediction is a prediction that consider more than one target variable in a real-life problems like cervical cancer simultaneously instead of the concentration of most researchers on supervised learning that has to do with prediction of a single target variable. The framework for multiple target variables has significant effect for categorization and evaluation that a single target variable framework cannot take care of. In the findings in the course of this study we did not come across the use of multi-target regression technique for predictive performance measure on cervical cancer dataset that predict all the target variables simultaneously. In this study, we adopt the problem transformation approach using multi-target classifiers to transform a binary classification task into a regression task. The predictive performance measures in supervised learning for multi-target classification task employ in this study is evaluated using exact match, hamming loss, hamming score, ZeroOne loss and accuracy per label. The findings of this study shows that the multi-target classifier (Bayesian classifier chains) using decision stump (base classifier) gives the highest predictive performance measures on hamming score, exact match, ZeroOne loss and accuracy per label compared to the multi-target classifier (classifier chains and class relevant) using J48 and random forest (base classifier) using 10 folds cross-validation and training and testing evaluation option. In conclusion, this study support the assertion made by some researchers that decision tree and random forest are powerful techniques for prediction.

Key words: Cervical cancer, multi-target variable, single target variable, multi-target regression, classifier, performance measures

INTRODUCTION

The word cancer is an umbrella term that refers to a class of diseases characterised by an abnormal growth of cells that divide and invade healthy body cells. The cancer cells that invade surrounding body tissue-cells and spread to other areas in the body then develops into cancer when the body's normal control mechanism stops functioning. Old body cells do not die out instead they grow out of control, accumulate and form a mass of a tissue called a tumour. Some cancers however, do not form tumours. About 70% of deaths due to cancer occurs in the 3rd world countries (Rahib *et al*., 2014; Shaha *et al*., 2008). Stewart and Kleihus (2003) records that there are different types of cancers and the various types of cancers are named according to the body part that is affected by the cancer cells. The various types of cancer include the following breast cancer, lung cancer, prostate cancer, colorectal cancer, skin cancer, cervical cancer, pancreatic cancer, ovarian cancer, gastric cancer, thyroid cancer to mention but a few.

Basically, there are two types of prevalent cancers in Swaziland namely the prostate cancer which affects men only and cervical cancer which affects women. Both cancers are curable if they are screened and detected at an early stage. Prostate cancer begins in the cells of the prostate gland (American Cancer Society, 2014), the prostate gland is the gland that is responsible for secreting some fluid that combines with sperm to form semen in men (Banerjee and Kalvani, 2016). There are several risk factors that are associated with the development of prostate cancer in Swazi men. According to Hayes and Bornman the risk of developing prostate cancer increases with age. Several studies have suggested that it is rare for men under the age of 40 to develop prostate cancer; however prostate cancer risks rapidly escalate age after the age of 50. Statistically, about 6 in 10 cases of prostate cancer are found in men over the age of 65 (Banerjee and Kalvani, 2016). Other prostate cancer risk factors are inflammation of the prostate gland (prostatitis), poor diet, sexually transmitted diseases such as gonorrhoea or chlamydia and having done vasectomy (American Cancer Society, 2014). Moreover, cervical cancer is the cancer that starts in the cervix usually developing from the mucosa of the surface of the cervix (Balogun *et al*., 2012). It is a common

Corresponding Author: S.G. Fashoto, Department of Computer Science, University of Swaziland, Kwaluseni, Swaziland

gynaecological cancer in women between the ages 25-30 (Balogun *et al*., 2012) and the most malignancy in incidence of mortality in Swazi women. According to the stop cervical cancer annual report of 2016 and other research studies, the Human Papilloma Virus (HPV) has been emphasized as the major cause of cervical cancer. There are at least 13 HPV types that cause cervical cancer of which HPV 16 and HPV 18 have been diagnosed as the major causes of cervical cancer. Constant and persistent infection with high risk HPV types sparks alteration in the cells of the cervix, the cells are known as precursor lesions or Cervical Intraepithelial Neoplasia (CIN). The prevalence of cervical cancer in Swazi women is made worse by Human Immunodeficiency Virus (HIV), the Southern African Litigation Centre (SALC) of 2012 report states that women living with HIV are 5 times likely to be infected with HPV thus increasing their susceptibility to develop cervical cancer. Balogun *et al*. (2012) pinpoints that apart from constant infection with HPV, early sexual debut, the use of birth control pills having contracted a Sexually Transmitted Infection (STI) having multiple sex partners and having a partner who has multiple sex partners exposes women to high chances of developing cervical cancer.

This study will focus on cervical cancer in Swazi women. The National health policy of Swaziland has noted that cervical cancer accounts for a percentage of 43.1% of all cancers among women thus cervical cancer is rendered as the most fatal cancer in Swazi women. According to Malambo , cervical cancer has greatly been neglected within the sector of Swazi public health. This glaring neglect is made crystal clear when observing available screening and treatment options for cervical cancer as compared with those of other diseases such as HIV/AIDS and tuberculosis in particular. Cervical cancer has been labelled as a 'silent' killer amongst Swazi women solely because of the low level of awareness about women's health in rural areas in particular where a large population of Swazi women resides. Also, Papanicolau (Pap) smear which is a procedure that entails the observation of cervical cells under a microscope, so as to ascertain if there are any cancer cells in the cervix is a procedure that has been met by active resistance and fear among Swazi women (Okonda *et al*., 2009). According to Lowy (2010), a Pap smear may alter a woman's sense of body and self, thus instilling fear, apprehension and in some case anxiety. Cervical screening entails stripping off your clothes and allowing a 'stranger' (a medical practitioner) to access your most intimate body part, it is for that reason, therefore, that cervical screening has been described as an 'invasion of privacy' by Armstrong *et al*. (2011) and McKie (1995) which therefore heightens the sense of vulnerability amongst Swazi women.

Literature review

Related works on cervical cancer: Cervical cancer is the most common cancer among Swazi women (Human Papillomavirus and Related Diseases in Swaziland, Summary Report, 27th July 2017). Most of the women rarely go for cervical cancer screening in developing countries such as Swaziland. According the Human Papillomavirus and related diseases in Swaziland summary report of 2017, there are about 223 new cervical cancer cases and 118 deaths a year.

A study conducted on 655 women in Swaziland Ginindza *et al*. (2017) showed an overall weighted high risk Human Papillomavirus (hr-HPV) prevalence of 46.2%. The most infected age group were women below 30 years with the prevalence decreasing with ages above 30 years. Swaziland also has a high HIV/AIDS prevalence and it was found that 24.4% of the women infected with hr-HPV are co-infected with HIV/AIDS.

Thangavel *et al*. (2006) discusses four categories of cervical cancer risk factors. The first category consists of the factors with which high association with cervical cancer has been proven. These include white discharge, hip pain, smelly vaginal discharge, early marriage, early sex and multiple sexual partners. In the second category are the factors which association with cervical cancer is likely such as malnutrition, male sexual behaviours among others. The third category is those associated with increased cervical cancer rate but can be influenced such as psychological factors. Lastly, the fourth category is the factors associated with increased cervical cancer rate which cannot be influenced, i.e., age, family and history.

There are two screening approaches used for diagnosing cervical cancer (Jusman *et al*., 2014). The first one is based on cellular level such as the common Pap smear and the other is based on tissue level such as visual inspection after applying Lugol's Iodine (VILI). These methods are highly dependent on the skill of experts. Computational tools were then developed to reduce the disadvantages that come with the two methods. These tools help the doctor with decision making.

By Jusman *et al*. (2014), discusses different computational intelligence techniques that can be used to develop diagnostic tools. These methods include Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), Logistic Regression and Decision Tree (DT). From the analysis of results, it was concluded that cervical precancerous data such as cytology, FISH, electromagnetic spectra, cervicography, colposcopy and HSDI can be used for screening purposes using the intelligent systems. Cellular level data produced better

results as compared to tissue level data and the ANN approach was the best performing among all the compared methods.

Jaganathan and Easmi used the k-means algorithm to come up with a tool that can help in diagnosing cervical cancer.

In this study, we evaluate the predictive performance measure parameters in multi-target classifiers and base classifiers on cervical cancer dataset for classification using Multi-Label extension of WEKA (MEKA) based on the use of the default training and testing split of the dataset.

Classifier is a rule for grouping object into pre-set cluster or class based on its attributes. Several classifiers are produced through influencing the training set (such as in boosting or bagging), influencing the input attributes, influencing the output targets or injecting randomness in the learning algorithm (Dzeroski and Zenko, 2004).

Currently, interest is more on non-parametric classifiers such as Support Vector Machine (SVM) and Classification Tree (CT) given that the flexibility of the required data and the exceptional experimental performance. Additionally, current studies presumed that an ensemble of classifiers may be an efficient method to increase classification accuracy and also limit prediction variation of single classifiers (Valentini and Dietterich, 2000).

Ensemble techniques are meta-classifiers which merge numerous machine learning techniques into a predictive model, so as to reduce variance (bagging), bias (boosting) or increase predictions (stacking).

MATERIALS AND METHODS

Multi-target prediction: Multi-target prediction is a prediction that consider more than one target variable in a real-life problems like cervical cancer simultaneously instead of the concentration of most researchers on supervised learning of classification model that has to do with prediction of a single target variable.

Machine learning classifiers normally aid single target variable. In term of regression model, the target is the actual value while in classification model the target can be one, two or more value. In classification models, the setback with multiple target variables is called multi-label classification. Within multi-target learning, a data instance is connected with multiple target variables such that each variable takes a number of values.

Types of regression model: There are three types of regression model used for classification in machine learning and they are as follows:

Simple regression model is a regression model that tries to use a suitable linear regression model on an independent variable. Multiple regression model is a regression model that tries to predict a dependent variable which is based on the value of two or more independent variables.

Multi-target regression is a regression model used on two or more dependent variables. If the target variables are categorical, it can be described as multi-label or multi-target classification, but if the target variables are numeric, subsequently multi-target regression is the name commonly used.

As in multi-label classification, there is a common problem transformation method that transforms the multi-target regression problem into multiple single-target regression problems. In this case, we consider each numeric target separately and train a single-target regressor for each of them. However, this local approach suffers from similar problems as the problem transformation approaches to multi-label classification. The single target models do not consider the inter-correlations of the target variables. The task of simultaneous prediction of all target variables at the same time (the global approach) has been considered in the batch setting by Struyf and D_eroski.

The most common approach employ using multi-label classification model is the problem transformation instead of the algorithmic adaptation.

The multi-target predictive supervised learning performance evaluation measures can be categorised into two and they are example-based and label-based. The Example-based measures rely on the mean difference of the real and predicted groups of labels over the entire set of data examples from the evaluation set while the label-based measures evaluate the performance for each label separately and find the mean of the performance over all labels. The classification-based evaluation predictive performance measures are considered in this study while the ranking-based are not. The predictive performance measures in supervised learning for single target based is conventionally evaluated using accuracy, f-measure, area under the receiver operating cost Curve (AUC) and so on while that of the multi-target based for classification is evaluated using Hamming loss, hamming score, rank loss, average precision, ZeroOne, One error and so on.

Data collection: In this study, 858 datasets of cervical cancer was obtained from University of California Irvine (UCI) repository as a prototype for the cervical cancer dataset that will be collected from the Ministry of Health, Swaziland. The cervical cancer dataset contains three target variables and each target variable has two class

Table 1: Attributes of cervical cancer

Attributes	Data type
Age	Int
Number of sexual partners	Int
First sexual intercourse(age)	Int
Num of pregnancies	Int
Smokes	Bool
Smokes (years)	Bool
Smokes (packs/year)	Bool
Hormonal contraceptives	Bool
Hormonal contraceptives (years)	Int
IUD	Bool
STDs	Bool
STDs (number)	Int
STDs: Condylomatosis	Bool
STDs: Cervical condylomatosis	Bool
STDs: Vaginal condylomatosis	Bool
STDs: Vulvo-perineal condylomatosis	Bool
STDs: Syphilis	Bool
STDs: Pelvic inflammatory disease	Bool
STDs: Genital herpes	Bool
STDs: Molluscum contagiosum	Bool
STDs: AIDS	Bool
STDs: HIV	Bool
STDs: Hepatitis B	Bool
STDs: HPV	Bool
STDs: Number of diagnosis	Int
STDs: Time, since, first diagnosis	Int
STDs: Time, since, last diagnosis	Int
Dx:cancer	Bool
Dx: CIN	Bool
Dx: HPV	Bool
Dx:	Bool
Hinselmann	Bool
Schiller	Bool (Target variable)
Cytology	Bool (Target variable)
Biopsy	Bool (Target variable)

labels and 33 nominal attributes. The 3 targets variables are schiller, cytology and biopsy. The attributes and the data types of the cervical cancer are presented in Table 1.

The cervical cancer dataset is fed into pre-processing phase using replace missing values method to handle the missing values. The reorder filter phase is applied to order the 3 target variables to look different from the order of the single target variable in the dataset. In the third phase, two different evaluation methods were used and they are training/testing and 10 fold cross-validation. The training/testing method is divided into two parts using the training and the testing split. The 67% was used for training the dataset and 33% for testing the dataset. The base classifiers phase use the J48 and 10 fold cross-validation while the multi-target classifiers phase use Bayesian Classier Chains (BCC), Classifier Chains (CC) and the Class Relevant (CR). The multi-target classifiers and base classifiers produce the predictive performance measures of the cervical cancer as shown in Fig. 1.

In this study, we adopt the problem transformation approach using multi-target classifiers to transform a

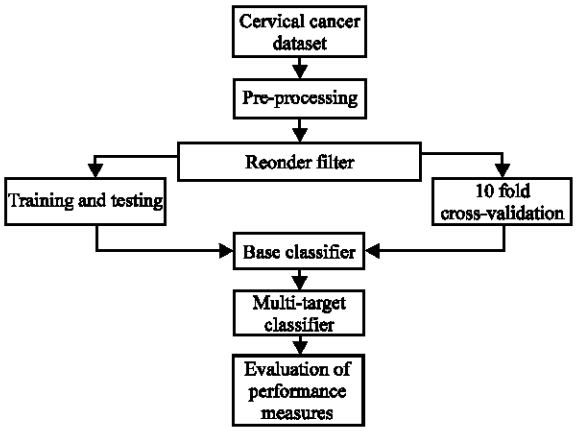

Fig. 1: The multi-target prediction classifiers

binary classification task into a regression task which is the simplest scenario. The multi-target variables have a binary target with labels yes and no and the binary target is considered a numeric target. The numeric target assigned a numeric value of one to the binary label yes and zero to the binary label no. The description of the simulation implemented is represented in Fig. 1.

Evaluation of the predictive performance measures on classifier models: In the predictive performance measures, we employ the recently commonly used parameters by researchers on multi-label algorithms (Gibaja and Ventura, 2015; Osojnik *et al*., 2011). The predictive performance measures are as follows.

Accuracy is the prediction of $\tilde{y}$ example-based on a real labelset y and the Jaccard Similarity Coefficient (JSC) between them:

$$JSC = \frac{|\hat{y} \cap y|}{\hat{y} \cup y|}$$

$$Accuracy = \frac{1}{p}\sum_{i}^{p} i = 1 \frac{|\hat{y}_i \cap y_i|}{\hat{y}_i \cup y_i|}$$

An increase in accuracy, improve the predictive performance measures while a decrease in accuracy, lower the predictive performance measures hamming loss is used to measure how many times an example-based and label-based is misclassified:

$$Hamming\ loss = \frac{1}{p}\sum_{i}^{p} i = 1 \frac{1}{Q}|\hat{y}_i \Delta y_i|$$

$\tilde{y}_i \Delta y_i$ is the symmetric difference of the sets $\tilde{y}$ and y. Hamming score = 1-Hamming loss.

For comparison purposes, the predictive performance measures parameters used in evaluating the three multi-target classifiers (Bayesian classifier chains, classifier chains and class relevant) and base classifiers (J48 and Random forest) include Hamming loss, hamming score, rank loss, average precision, ZeroOne loss, One error, F1 macro, F1 micro, Rank loss and so, on. Besides, the ranking method was used to single out the optimal model among meta-classifiers and the single classifier.

However, in this study hamming score, exact match, Hamming loss, ZeroOne loss and accuracy per label were used for evaluating the predictive performance measures of the multi-target classifiers.

RESULTS AND DISCUSSION

The proposed evaluation of the meta-classifiers and single classifier on cervical cancer classification system was simulated using multi-Label extension of WEKA (MEKA 1.9.0). The multi-target classifiers used in this paper are the Bayesian Classifier Chains (BCC), Classifier Chains (CC) and Class Relevant (CR) while the base classifiers employ is the J48 and random forest. For comparison purposes, multi-target and base classifiers are based on five predictive performance measures parameters namely hamming score, exact match, hamming loss, ZeroOne loss and accuracy per label for the evaluation.

Comparison of multi-target classifiers and base classifiers: The multi-target classifier (BCC) using Decision Stump base classifier gives the highest predictive performance measures on hamming score, exact match, ZeroOne loss and accuracy per label compared to the multi-target classifier (CC and CR) using J48 and random forest on Table 2.

The multi-target classifier (BCC) using Decision Stump base classifier gives the highest predictive performance measures on hamming score, exact match, ZeroOne loss and accuracy per label compared to the multi-target classifier (CC and CR) using J48 and random forest on Table 3.

The results for the single target variable and the optimal multi-target variables prediction for performance measures are as follows.

Using 10 folds cross-validation with Biopsy as the target variable for the cervical cancer dataset, the predictive performance accuracy is 52.9% on decision stump while the predictive performance accuracy is 66.4% using training and testing while that of the multi-target variables prediction for label 2 (Biopsy) performance accuracy measure is 95.3% for 10 folds cross-validation and 96.6% for training and testing.

Using 10 folds cross-validation with Schiller as the target variable for the cervical cancer dataset, the predictive performance accuracy is 64.8% on decision stump while the predictive performance accuracy is 63.6%

Table 2: Train/test split

Multi-target classifier/base classifier	Hamming score	Exact match	Hamming loss	ZeroOne loss	Accuracy per label
BCC					
J48	0.951	0.897	0.049	0.103	0.945 (0) 0.955 (1) 0.952 (2)
Random forest	0.958	0.911	0.042	0.089	0.949 (0) 0.962 (1) 0.962 (2)
Decision stump	0.958	0.911	0.042	0.089	0.945 (0) 0.962 (1) 0.966 (2)
CR					
J48	0.952	0.897	0.048	0.099	0.945 (0) 0.945 (1) 0.966 (2)
Random forest	0.952	0.901	0.048	0.099	0.945 (0) 0.945 (1) 0.966 (2)
Decision stump	0.958	0.911	0.042	0.089	0.945 (0) 0.962 (1) 0.966 (2)
CC					
J48	0.953	0.897	0.047	0.103	0.945 (0) 0.955 (1) 0.959 (2)
Random forest	0.953	0.897	0.047	0.103	0.945 (0) 0.955 (1) 0.959 (2)
Decision stump	0.958	0.911	0.042	0.089	0.949 (0) 0.955 (1) 0.959 (2)

Table 3: The 10 fold cross-validation

Multi-target classifier/base classifier	Hamming score	Exact match	Hamming loss	zeroOne loss	Accuracy per label
BCC					
J48	0.949	0.899	0.051	0.101	0.952 (0)
					0.948 (1)
					0.948 (2)
Random forest	0.949	0.899	0.051	0.101	0.952 (0)
					0.948 (1)
					0.948 (2)
Decision stump	0.951	0.903	0.049	0.097	0.952 (0)
					0.949 (1)
					0.953 (2)
CR					
J48	0.951	0.901	0.049	0.099	0.952 (0)
					0.948 (1)
					0.952 (2)
Random forest	0.951	0.901	0.049	0.099	0.952 (0)
					0.948 (1)
					0.952 (2)
Decision stump	0.951	0.903	0.049	0.097	0.952 (0)
					0.949 (1)
					0.953 (2)
CC					
J48	0.949	0.899	0.051	0.101	0.952 (0)
					0.948 (1)
					0.949 (2)
Random forest	0.949	0.899	0.051	0.101	0.952 (0)
					0.948 (1)
					0.949 (2)
Decision stump	0.951	0.903	0.049	0.097	0.952 (0)
					0.949 (1)
					0.953 (2)

using training and testing while that of the multi-target variable prediction for label 1 (Schiller) performance accuracy measure is 94.9% for 10folds cross-validation and 96.2% for training and testing.

Using 10 folds cross-validation with Citology as the target variable for the cervical cancer dataset, the predictive performance accuracy is 16.2% on decision stump while the predictive performance accuracy is 2.6% using training and testing while that of the multi-target variable prediction for label 0 (cytology) performance accuracy measure is 95.2% for 10 folds cross-validation and 94.5% for training and testing.

CONCLUSION

Reorder filter and pre-processing is an important phase to the improvement of predictive performance measures, especially, for multiple target variables and missing values in a cervical cancer dataset. This study presents a comparison between different types of multi-target classifiers and different types of base classifiers on a cervical cancer dataset simulation for predictive performance measures evaluation. The prediction accuracy of the multi-target classifiers and the base classifiers models depends on the evaluation options used in MEKA (10 fold cross-validation or percentage split 66%) on the dataset for the predictive performance measures evaluation. The first findings show that the multi-target classifier (BCC) using decision stump base classifier gives the highest predictive performance measures on hamming score, exact match, ZeroOne loss and accuracy per label compared to the multi-target classifier (CC and CR) using J48 and random forest.

The second findings show that the multi-target classifier (BCC) using decision stump base classifier gives the highest predictive performance measures on hamming score, exact match, ZeroOne loss and accuracy per label compared to the multi-target classifier (CC and CR) using J48 and random forest on. These two findings show that the multi-target variables prediction outperforms the single target variable prediction using either the training and testing or the 10 folds cross-validation evaluation option.

REFERENCES

American Cancer Society, 2014. Colorectal cancer facts and figures 2014-2016. American Cancer Society, USA., pp: 1-32. http://www.cancer.org/acs/groups/content/documents/document/acspc-042280.pdf.

Armstrong, N., V. James and M. Dixon-Woods, 2011. The role of primary care professionals in women's experiences of cervical cancer screening: A qualitative study. Family Pract., 29: 462-466.

Balogun, M.R., O.O. Odukoya, M.A. Oyediran and P.I. Ujomu, 2012. Cervical cancer awareness and preventive practices: A challenge for female urban slum dwellers in Lagos, Nigeria. Afr. J. Reprod. Health, 16: 75-82.

Banerjee, S. and A. Kaviani, 2016. Worldwide prostate cancer epidemiology: Differences between regions, races and awareness programs. Intl. J. Clin. Exp. Med. Sci., 2: 1-16.

Dzeroski, S. and B. Zenko, 2004. Is combining classifiers with stacking better than selecting the best one?. Mach. Learn., 54: 255-273.

Gibaja, E. and S. Ventura, 2015. A tutorial on multilabel learning. ACM. Comput. Surv., 47: 52-52.

Ginindza, T.G., X. Dlamini, M. Almonte, R. Herrero and P.E. Jolly *et al*., 2017. Prevalence of and associated risk factors for high risk human papillomavirus among sexually active women, Swaziland. PloS One, 12: e0170189-1-e0170189-18.

Jusman, Y., S.C. Ng, A. Osman and N. Azuan, 2014. Intelligent screening systems for cervical cancer. Sci. World J., 2014: 1-15.

Lowy, I., 2010. Cancer, women and public health: The history of screening for cervical cancer. Hist. Cienc. Saude Manguinhos, 17: 53-67.

McKie, L., 1995. The art of surveillance or reasonable prevention? The case of cervical screening. Sociol. Health Illness, 17: 441-457.

Okonda, S., C. Wright and P. Michelow, 2009. The status of cervical cytology in Swaziland, Southern Africa: A descriptive study. Cyto J., Vol. 6, 10.4103/1742-6413. 54916.

Osojnik, A., P. Panov and S. Dzeroski, 2017. Multi-label classification via multi-target regression on data streams. Mach. Learn., 106: 745-770.

Rahib, L., B.D. Smith, R. Aizenberg, A.B. Rosenzweig and J.M. Fleshman *et al*., 2014. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver and pancreas cancers in the United States. Cancer Res., 74: 2913-2921.

Shaha, M., C.L. Cox, K. Talman and D. Kelly, 2008. Uncertainty in breast, prostate and colorectal cancer: Implications for supportive care. J. Nurs. Scholarship, 40: 60-67.

Stewart, B.W. and P. Kleihues, 2003. EDS World Cancer Report. IARC Press, Lyon, pp: 22-47.

Thangavel, K., P.P. Jaganathan and P.O. Easmi, 2006. Data mining approach to cervical cancer patients analysis using clustering technique. Asia J. Inform. Technol., 5: 413-417.

Valentini, G. and T.G. Dietterich, 2004. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. J. Mach. Learn. Res., 5: 725-775.