Colon Cancer Classification and Patients' Survival Detection Using Support Vector Machine Kernels

Kamoru Jimoh[1], Yakub Kayode Saheed[2], and Maruf O Alimi[3]
[1]Department of Physical Science, Al-Hikmah University, Ilorin, Nigeria
[2,3]Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria.
E-mails: jimkaminsha@alhikmah.edu.ng[1], yksaheed@alhikmah.edu.ng[2], moalimi@alhikmah.edu.ng[3]
Phone: +2349038040012, +2348142683364, +2348171242008

ABSTRACT

Colon cancer (CC) is the third most common cancer in the world. Due to the high rate of CC and the benefits of data mining to predict its survival rate, the aim of this study is to investigate the performance of machine learning algorithms; Support Vector machine Poly Kernel (SVM-PK) and Support Vector Machine-Gaussian kernel (SVM-GK), to predict the survival of CC patients. This study employed CC data obtained from the University of California Irvine (UCI) machine learning repository. We used the SVM-PK and SVM-GK methods in order to predict the survival of patients with CC. The experimental analysis was performed in Matrix laboratory (MATLAB 2015) environment. The performance of the two algorithms was evaluated using confusion matrix. The experimental results obtained gave prediction accuracy of 99%, sensitivity of 99%, and 98% specificity was obtained for SVM-PK while SVM-GK gave accuracy of 99%, sensitivity of 99% and specificity of 96%. Our results outperformed the related state-of-art in terms of accuracy, sensitivity and specificity of CC survival prediction.

Keywords: Colon cancer, Support Vector Machine Poly kernel, Support Vector Machine Gaussian Kernel, Data mining

iSTEAMS Conference Proceedings Paper Citation Format
Waidor. K. Tamaramiebi & Akpojaro Jackson (2018): The Use of Classification Algorithm for Students Academic Performance Forecasting. Proceedings of the 14th iSTEAMS International Multidisciplinary Conference, AlHikmah University, Ilorin, Nigeria, Vol. 14, Pp 159-164

1. INTRODUCTION

Colon cancer (CC) is the third most common cancer in the world (Setareh, Zahiri, Zare, and Abbasi, 2018; Reza, Mitra, and Mohammad, 2015). Cancers of the colon and rectum are of two of the most common types worldwide. Early diagnosis and treatment can greatly improve the chances of survivability (Reda, Ankit and Alok, 2013). In the past decade, Computer science, Statistics and medical fields have been involved in providing diagnosis of various human diseases. Information generated from patients to medical personnel in biomedical prognosis and diagnosis may include redundant, irrelevant, and interrelated symptoms most often in the case whereby a patient suffers from more than one type of disease of the same category. Hence, it becomes a serious challenge for physician to diagnose perfectly. Early detection and accurate prediction is achievable by medical personnel using data mining technique in health care industry (Himanshu and Rizvi, 2017).

Data mining (DM) also known as knowledge discovery in databases (KDD), is a process that aims to discover relationships between items and hidden information from large datasets (Abdulsalam et al., 2017; Tseng, 2015). DM has been used recently and successfully in bioinformatics (Ayomikun,Oladele and Saheed, 2018; Jimoh, Yusuf, and Saheed, 2018), electric load forecasting (Hambali et al., 2017) and educational data mining (Abdulsalam et al., 2017). The techniques in DM have contributed immensely in transforming large data into specific and more relevant information for knowledge discovery and prediction purpose (Agrawal and Chopde, 2016). Accuracy is the vital thing to be considered during estimation over colon data. DM techniques can be grouped as association, classification and clustering. Association works on the basis of correlation, classification helps in categorizing and locating accurately, and clustering is the unsupervised learning ability that is able to discover hidden patterns of dataset. The objective of this paper is to investigate the classification performance of Support Vector Machine Kernels (SVM-K) which are Support Vector Machine-Poly kernel (SVM-PK) and Support Vector Machine-Gaussian kernel (SVM-GK), to predict the outcome of CC patients.

This paper is organized as follows. Section 2 presents the related work. In section 3 the materials and methodology is presented. Section 4 highlights the results and discussion and section 5 concludes the paper.

2. RELATED WORK

Setareh et al. (2018) proposed a survey on two widely used machine learning algorithms, Bagging and Support Vector Machines (SVM), to predict the outcome of colon cancer patients. The Weka software ver 3.6.10 was used for data analysis. The performance of two algorithms was determined using the confusion matrix. The accuracy, specificity, and sensitivity of the SVM was 84.48%, 81%, and 87%, and the accuracy, specificity, and sensitivity of Bagging was 83.95%, 78%, and 88%, respectively.

The authors in (Reza, Mitra, and Mohammad, 2015) presented a rule based data mining classification techniques for colon cancer survivability. They employed Trees Random Forest (TRF), AdaBoost (AD), RBF Network (RBF-N), and Multilayer Perceptron (MLP) machine learning techniques with 10-cross fold technique with the proposed model for the prediction of colon cancer survival. The performance of machine learning techniques were evaluated with accuracy, precision, sensitivity, specificity, and area under ROC curve. Results obtained showed that TRF which is a rule based classification model has the highest level of accuracy.

Kaladhar et al. (2013) presented a classification and clustering technique for colon cancer survivability using Logistics, Ibk, Kstar, NNge, Naïve Bayes, ADTree, and Random Forest Algorithms. The experiment was performed in weka data mining suite. The results obtained gave an accuracy of 97.22% for Naïve Bayes, which was the highest accuracy for all algorithms used in this study.

The researchers (Reda, Ankit and Alok, 2013) in developing survival prediction model for colon cancer, carefully designed a preprocessing step using synthetic minority over-sampling technique (SMOTE) to balance the survival and non survival classes. In this experiment, ensemble voting of the three of the top performing classifiers was found to result in the best prediction performance in terms of prediction accuracy and area under the ROC curve.
As seen from the findings of the literature, it is observed that still there is a big demand to have better and higher accuracy. Also, the feature extraction stage before classification is an important issue that has not been well addressed in the literature. Hence, there is need for more research for improving the colon cancer survivability detection towards the improvement of the classification technique.

3. MATERIALS AND METHODS

The techniques and methods required to achieve the stipulated objectives are explained in the sub-categories below.
Requirement Phase

Data set acquisition
A portion of the colon Dataset was extracted from the UCI Machine Learning repository for the purpose of feeding the developed model.

Dataset Description
The experiment for training and testing of the proposed intelligent approach for colon cancer prediction survivability was applied by using a real dataset. These datasets contain a standard set of data to be audited and the datasets include a wide variety of intrusion types simulated in a network environment.

Dataset Pre-processing
Pre-processing of original data set is an important phase to make it as an appropriate input for classification phase. The main objective of preprocessing phase is to reduce ambiguity and provide accurate information to detection engine. The preprocessing phase cleans the data by grouping, labeling and it handles the missing or incomplete dataset.

Data set pre-processing
Getting rid of errors and outliers that are present in the data are parts of pre-processing task that was done to make the data suitable for modelling. This part focused to filter out every inconsistent set of the data set thereby enhancing a smooth operation on the dataset for better result optimization.

Feature extraction

In order to perform feature extraction, the Principal Component Analysis (PCA) algorithm would be employed to extract features with high components.

Feature classification.

After feature classification the main analysis of classification was performed with SVM-PK and SVM-GK. The SVM-PK and SVM-GK formulate a machine learning technique or model that would assist to predict the colon cancer.

Design Phase

The Design phase was implemented using MATLAB Guide and involved the loading of the dataset from an excel file sheet on a designed and developed interactive graphical user interface that incorporated all the activities of the simulation in a friendly environment.

Coding Phase

The Coding phase integrate the Graphical User Interface (GUI) with all the necessary developed data mining functions for performing task and operation in the developmental process thereby also enhancing communication between the User Interface and the Database Knowledge Discovery.

Model Evaluation

Various statistical performance metrics were used to measure the effectiveness of the model in terms of the prediction power and accuracy. The metrics are true positive rate, false negative rate, false positive rate, classifier accuracy, precision and recall. To get this value the supplied dataset will be divided into two.

Training Set

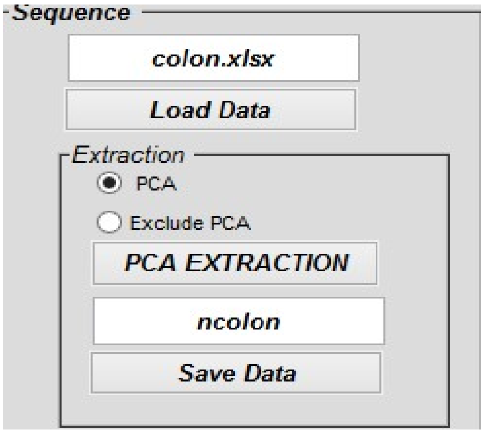
The training set occupied 75% of the dataset. The 75% dataset will help to train the model thereby creating an experiment knowledge base.

Testing set

The testing set occupied a 25% portion of the dataset, this will serve as a testing dataset for the trained model so as to check and validate the model.

4. RESULTS AND DISCUSSION

The colon data set is first loaded into the proposed model. The figure 1 showed the dataset loaded into the system. The colon data loaded into the model is then extracted using PCA as depicted in figure 1 and 2. The data extracted has 25192 observations and 42 attributes.



Sequence

colon.xlsx

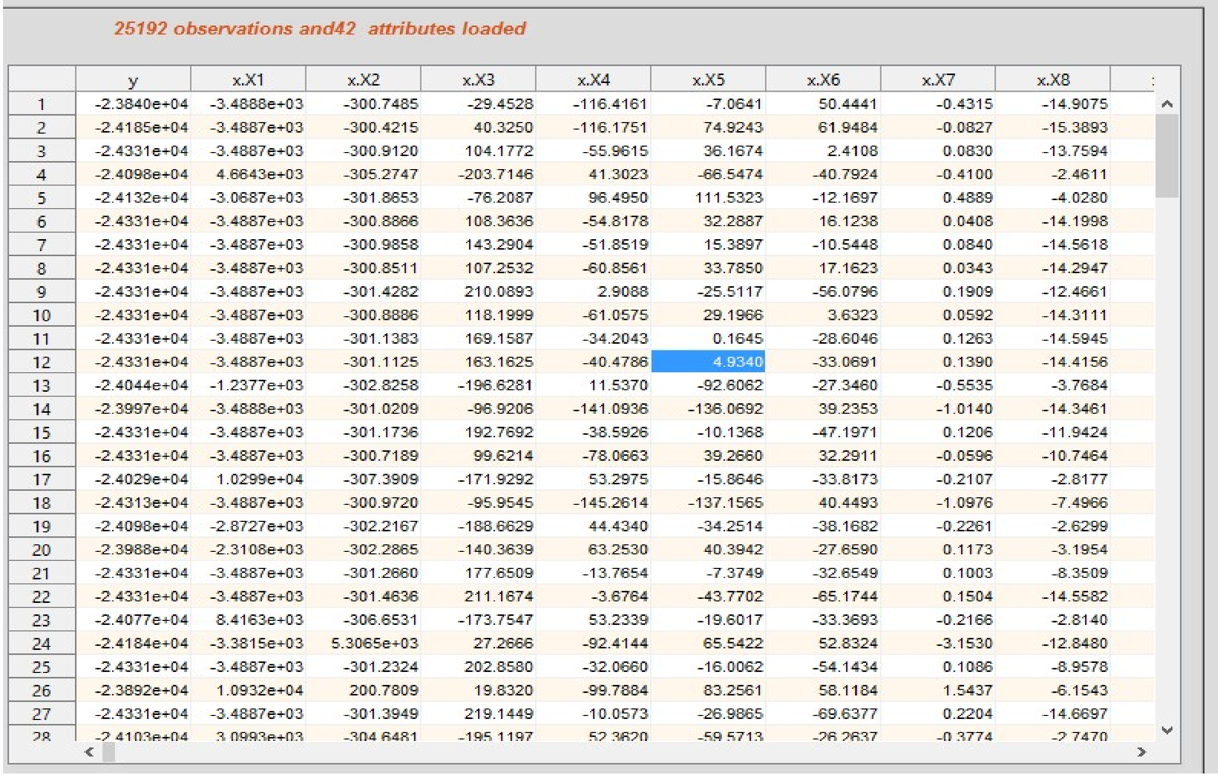
Extraction

● PCA

○ Exclude PCA

ncolon

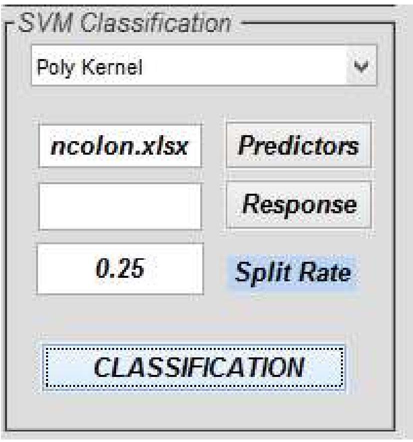
Fig.1. PCA extraction phase



25192 observations and42 attributes loaded

	y	x.X1	x.X2	x.X3	x.X4	x.X5	x.X6	x.X7	x.X8
1	-2.3840e+04	-3.4888e+03	-300.7485	-29.4528	-116.4161	-7.0641	50.4441	-0.4315	-14.9075
2	-2.4185e+04	-3.4887e+03	-300.4215	40.3250	-116.1751	74.9243	61.9484	-0.0827	-15.3893
3	-2.4331e+04	-3.4887e+03	-300.9120	104.1772	-55.9615	36.1674	2.4108	0.0830	-13.7594
4	-2.4098e+04	4.6643e+03	-305.2747	-203.7146	41.3023	-66.5474	-40.7924	-0.4100	-2.4611
5	-2.4132e+04	-3.0687e+03	-301.8653	-76.2087	96.4950	111.5323	-12.1697	0.4889	-4.0280
6	-2.4331e+04	-3.4887e+03	-300.8866	108.3636	-54.8178	32.2887	16.1238	0.0408	-14.1998
7	-2.4331e+04	-3.4887e+03	-300.9858	143.2904	-51.8519	15.3897	-10.5448	0.0840	-14.5618
8	-2.4331e+04	-3.4887e+03	-300.8511	107.2532	-60.8561	33.7850	17.1623	0.0343	-14.2947
9	-2.4331e+04	-3.4887e+03	-301.4282	210.0893	2.9088	-25.5117	-56.0796	0.1909	-12.4661
10	-2.4331e+04	-3.4887e+03	-300.8886	118.1999	-61.0575	29.1966	3.6323	0.0592	-14.3111
11	-2.4331e+04	-3.4887e+03	-301.1383	169.1587	-34.2043	0.1645	-28.6046	0.1263	-14.5945
12	-2.4331e+04	-3.4887e+03	-301.1125	163.1625	-40.4786	4.9340	-33.0691	0.1390	-14.4156
13	-2.4044e+04	-1.2377e+03	-302.8258	-196.6281	11.5370	-92.6062	-27.3460	-0.5535	-3.7684
14	-2.3997e+04	-3.4888e+03	-301.0209	-96.9206	-141.0936	-136.0692	39.2353	-1.0140	-14.3461
15	-2.4331e+04	-3.4887e+03	-301.1736	192.7692	-38.5926	-10.1368	-47.1971	0.1206	-11.9424
16	-2.4331e+04	-3.4887e+03	-300.7189	99.6214	-78.0663	39.2660	32.2911	-0.0596	-10.7464
17	-2.4029e+04	1.0299e+04	-307.3909	-171.9292	53.2975	-15.8646	-33.8173	-0.2107	-2.8177
18	-2.4313e+04	-3.4887e+03	-300.9720	-95.9545	-145.2614	-137.1565	40.4493	-1.0976	-7.4966
19	-2.4098e+04	-2.8727e+03	-302.2167	-188.6629	44.4340	-34.2514	-38.1682	-0.2261	-2.6299
20	-2.3988e+04	-2.3108e+03	-302.2865	-140.3639	63.2530	40.3942	-27.6590	0.1173	-3.1954
21	-2.4331e+04	-3.4887e+03	-301.2660	177.6509	-13.7654	-7.3749	-32.6549	0.1003	-8.3509
22	-2.4331e+04	-3.4887e+03	-301.4636	211.1674	-3.6764	-43.7702	-65.1744	0.1504	-14.5582
23	-2.4077e+04	8.4163e+03	-306.6531	-173.7547	53.2339	-19.6017	-33.3693	-0.2166	-2.8140
24	-2.4184e+04	-3.3815e+03	5.3065e+03	27.2666	-92.4144	65.5422	52.8324	-3.1530	-12.8480
25	-2.4331e+04	-3.4887e+03	-301.2324	202.8580	-32.0660	-16.0062	-54.1434	0.1086	-8.9578
26	-2.3892e+04	1.0932e+04	200.7809	19.8320	-99.7884	83.2561	58.1184	1.5437	-6.1543
27	-2.4331e+04	-3.4887e+03	-301.3949	219.1449	-10.0573	-26.9865	-69.6377	0.2204	-14.6697
28	-2.4103e+04	3.0993e+03	-304.6481	-195.1197	52.3620	-59.5713	-26.2637	-0.3774	-2.7470

Fig.2. Data extracted with PCA



SVM Classification

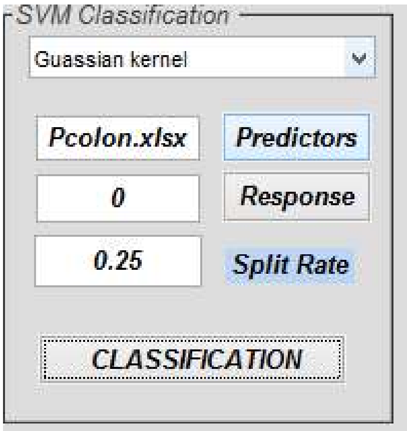
Poly Kernel

ncolon.xlsx

0.25

Fig.3. SVM-PK classification phase

The figure 3 and figure 4 illustrates the SVM-PK phase and SVM-GK with 25% hold out. This indicates that 75% of the dataset is used for training the model while 25% is used for testing.



SVM Classification

Guassian kernel

Pcolon.xlsx — Predictors

0 — Response

0.25 — Split Rate

Fig.4. SVM-GK classification phase

Table 1. The results of classifiers performance based on SVM-PK and SVM-GK

Classifiers/Metrics	Sensitivity (%)	Specificity (%)	F-Score (%)	Precision (%)	Recall (%)	Accuracy (%)
SVM-PK	99.1969	98.6035	98.9908	98.7855	99.1969	99.5681
SVM-GK	99.8513	96.594	98.4602	97.1073	99.8513	99.3331

As can be seen from Table 1, the SVM-PK has higher specificity, F-score, precision and accuracy, than the SVM-GK. However, the SVM-GK has better recall and sensitivity. The results obtained showed that SVM-PK performed better than the SVM-GK in classifying colon cancer.

Table 2. Comparison with the existing methods

Authors/Years	Classifiers/metrics	Accuracy (%)	Specificity (%)	Sensitivity (%)
Setareh et al. (2018)	SVM	84.	81	87
Setareh et al. (2018)	Bagging	83	78	88
Jimoh et al. (2018)	ICA	88	72	100
Jimoh et al. (2018)	KNN	77	36	74
Proposed method	SVM-PK	99	98	99
Proposed method	SVM-GK	99	96	99

The experimental result of our proposed methods was compared with the existing methods Jimoh et al. (2018) and Seterah et al. (2018). The results of our proposed method outperformed the existing methods in terms of accuracy, specificity and some sensitivity.

5. CONCLUSION

In this paper, we proposed classification schemes to construct models for survival prediction of CC patients. The CC data used was obtained from the UCI machine learning repository and SVM-PK and SVM-GK were used to predict the CC patients' survival. Experimental results obtained showed that the SVM-PK has higher specificity, F-score, precision and accuracy, than the SVM-GK. However, the SVM-GK has good recall and sensitivity. Thus, the proposed method performed better than the existing method. Future work includes exploring more techniques to deal with imbalanced data, redundancy and noise in the data. Additionally, the extraction time and the training time would be interesting variables in the future.

REFERENCES

[1] Agrawal, N.R., and P.A. Chopde. (2016). A Survey on Heart Disease Prediction Using Soft Computing. Int. J.Eng. Sci. Res., Vol.5, no.3, pp.582-587, 2016.

[2] Hambali, M. A., Saheed, Y. K, Gbolagade, M. D, Gaddafi M. (2017). Artificial Neural Network Approach For Electric Load Forecasting in Power Distribution Company. Volume 6 Issue 2 2017, 80-90.e-Academia Journal. http://journale-academiauitmt.uitm.edu.my Universiti Teknologi MARA Terengganu.

[3] Abdulsalam, S.O. Saheed, Y.K. Hambali, M.A. Salau-Ibrahim, T.T. Akinbowale, N.B. (2017). Student's Performance Analysis Using Decision Tree Algorithms. Annals. Computer Science Series. 15th Tome 1st Fasc. – 2017.

[4] Ayomikun, K.O., T.O. Oladele, & Y. K. Saheed, (2018). Comparative Evaluation of Linear Support Vector Machine and K-Nearest Neighbour Algorithm using Microarray Data On Leukemia Cancer Dataset, Afr. J. Comp. & ICT, Vol.11, No.2, pp. 1 - 10.

[5] Jimoh, R.G., R.M. Yusuf, Yusuf, O.O. Saheed, Y.K. (2018). Application of Dimensionality Reduction on Classification of Colon cancer Using ICA and K NN Algorithms. Annals. Computer Science Series. 16th Tome 1st Fasc. – 2018.

[6] Abdulsalam, S. O., Hambali, M. A. Salau-Ibrahim, T.T. Saheed, Y.K. Akinbowale, N.B. (2017). Knowledge Discovery From Educational Database Using Apriori Algorithm. GESJ: Computer Science and Telecommunications 2017.No.1(51).

[7] Tseng, W.T., Chiang, W.F., Liu, S.Y., Roan, J., and Lin, C.N., (2015). The application of data mining techniques to oral cancer prognosis. J Med Syst. 39(5):59, 2015. doi:10.1007/s10916-015-0241-3.

[8] Himanshu, S. and M. A. Rizvi., (2017). Prediction of Heart Disease using Machine LearningAlgorithms: A Survey. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 8.

[9] Reda Al-Bahrani, Ankit Agrawal, Alok Choudhary (2013). Colon cancer survival prediction using ensemble data mining on SEER data. 2013 IEEE International Conference on Big Data. Pp.9-16

[10] Reza Abbasi, Mitra Montazeri, Mohammad Zare (2015). A Rule Based Classification Model to Predict Colon Cancer Survival. IEEE International Conference on Deep Learning. Setareh S, Zahiri Esfahani M, Zare Bandamiri M, Raeesi A, Abbasi R. (2018). Using Data Mining for Survival Prediction in Patients with Colon Cancer. irje. 2018; 14 (1) :19-29.

[11] Kaladhar, D. S.V.G., Bharath Kumar Pottumuthu, Padmanabhuni V. Nageswara Rao, Varahalarao Vadlamudi, A.Krishna Chaitanya, R. Harikrishna Reddy (2013). The Elements of Statistical Learning in Colon Cancer Datasets: Data Mining, Inference and Prediction. Algorithms Research 2013, 2(1): 8-17