ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 10, April 2014

Evaluating the Accuracy of Ensemble Learning Approaches for Prediction on Recurrent Colorectal Cancer

Chi-Chang Chang*[1, 2], Wen-Chien Ting[3], Ting Teng[1], Che-Hsin Hsu[1]
[1] School of Medical Informatics, Chung Shan Medical University, *changintw@gmail.com*
[2] Information Technology Office of Chung Shan Medical University Hospital
[3] Cancer Center, Chung Shan Medical University Hospital
[4] Institute of Medicine, Chung Shan Medical University

Abstract*— this paper discusses ensemble learning approaches for the recurrence prediction of colorectal cancer. The standard ensemble modeling, where the prediction was based on a majority voting of the three prediction models: Multivariate Adaptive Regression Splines (MARS), C5.0 and Random Forests (RF). The medical records were from the medical center in Taiwan. Based on the result of this study, C5.0 approach is the most useful approach to the discovery of recurrence factors with colorectal cancer. Surgical Margins of the Primary Site and Pathologic Stage Group was most important and independent prognostic factor of recurrent colorectal cancer. To our knowledge, this is the first study using ensemble learning approaches for the analysis of risk factors for recurrent colorectal cancer, and results of this study will contribute to developing the clinical practice guideline for colorectal cancer.

***Index Terms*—Colorectal Cancer, Ensemble Model, MARS, C5.0, Random Forests**

I. INTRODUCTION

Colorectal cancer (CRC) is the leading cause of death from cancer in western countries [1] and has the highest rate of incidence and is the second most common cause of cancer death in both men and women in Taiwan [2]. The primary modality of treatment for colorectal cancer is surgery. However, although over two-thirds of patients with primary disease undergo potentially curative surgery where all gross tumor is removed, up to 50% of these will eventually die in the ensuing 5 years, the majority from local, regional or distant tumor recurrence. Adding to the problem is the difficulty in predicting the site of recurrence. This is, at the moment, difficult to do since primary colorectal cancers at different locations in the bowel may have different recurrence patterns [l, 3]. About 5 - 10% of CRC patients with stage I, 20% of stage II patients, and 35% of stage III patients relapse and die from cancer recurrence [2]. DNA aneuploidy, an accepted marker for CIN, is found in the majority of sporadic CRC and has been linked to poor prognosis [4-6]. One of the ways this can be achieved is by improving the ability to predict those that will recur, which could lead to more focused or intensive follow-up [7]. Confounding the situation, the optimal strategy to accurately detect recurrences at the earliest possible time is a highly debated concept in the current colorectal cancer literature. It is well known that most recurrences occur within 5 years [7]. To aid in this process, several patient and disease-related factors have been identified that can help better predict the risk of recurrence [7-9]. The existing literature on recurrent colorectal cancer reveals that factors are include:(1) Sex, (2) Primary Site, (3) Histology, (4) Behavior Code, (5) Grade, (6) Regional Lymph Nodes Examined, (7) Regional Lymph Nodes Positive, (8) Surgical Diagnostic and Staging Procedure at Other Facility, (9) Surgical Diagnostic and Staging Procedure at This Facility, (10) Clinical T, (11) Clinical N, (12) Clinical M, (13) Clinical Stage Group, (14) Pathologic T, (15) Pathologic N, (16) Pathologic M, (17) Pathologic Stage Group, (18) Surgical Margins of The Primary Site, (19) Sequence of RT and Surgery, (20) Sequence of Locoregional Therapy and Systemic Therapy, (21) Dose of RT, (22) Chemotherapy at Other Facility, (23) Chemotherapy at This Facility, (24) Vital Status, (25) Cause of Death. Within these broad categories, many studies have attempted to identify the plethora of factors that could enhance clinical management in the intervention of colorectal cancer. Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem [10, 11]. Against this background, the present study attempts to improve surveillance after treatment might lead to earlier detection of relapse, and precise assessment of recurrent status could improve outcome.

II. METHOD

Based on ensemble learning technique, such as classification has not been used to analyses the recurrence colorectal cancer [12, 13]. In this paper, we made an attempt to identify patterns from the database of the colorectal cancer patients using several advances ensemble learning techniques as follows.

A. Multivariate Adaptive Regression Splines (MARS) Algorithm

In general, implementing Multivariate Adaptive Regression Splines (MARS) involves a two step and the algorithm is implemented: Step 1.Start with the simplest model involving only the constant basis function. Step 2 is recursively applied until a model of pre-determined maximum complexity is derived. Finally, in the last stage, a

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 10, April 2014

pruning procedure is applied where those basis functions are removed that contribute least to the overall (least squares) goodness of fit. The MARS algorithm builds models of the form:

$$(x-t)_+ = \begin{cases} x-t, x>t \\ 0, \text{otherwise} \end{cases} \quad (1)$$

The MARS model for a dependent (outcome) variable y, and M terms, can be summarized in the following equation:

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m H_{km}(x_{v(k,m)}) \quad (2)$$

Function H is defined as:

$$H_{km}(x_{v(k,m)}) = \prod_{k=1}^{K} h_{km} \quad (3)$$

where $x_{v(k,m)}$ is the predictor in the k^{th} of the m^{th} attribute. During forward stepwise, a number of basis functions are added to the model according to a pre-determined maximum which should be considerably larger than the optimal. Specifically, MARS uses two-sided truncated functions of the form as basic functions for linear or nonlinear expansion, which approximates the relationships between the response and predictor variables. In the most general terms, least squares estimation is aimed at minimizing the sum of squared deviations of the observed values for the dependent variable from those predicted by the model [14].

B. C5.0 Algorithm

C5.0 classifier is a process for the classification and analysis of information hidden in large datasets, which retrieves useful information in the form of a decision tree, i.e., a structure tree flowchart like [15]. The algorithm use a greedy approach in which the decision trees are constructed in a top-down recursive divide and conquer manner on the basis of a training set employing an attribute selection measure. C5.0 makes some improvement on C4.5 such as, faster, more memory efficient, similar results by smaller decision trees, supports for more accuracy, weight different attributes and misclassification types, reduce noise [15, 16, 17]. Take calculating evaluation properties of A as an example, calculate information gain ratio $GainRatio(A)$, S represents a set of samples, p_i is the probability that an arbitrary sample belongs to B_i. Suppose that categorical attributes have n different values, which define n different classes B_i, (i = 1, ... ,n). Suppose S_i, is the number of samples in the class B. $Info(S)$ indicates the information entropy in the current sample. The calculation process is

$$Info(S) = \sum_{i=1}^{n} p_i \log(p_i) \quad (4)$$

Suppose attribute A has n different values $\{A_1, A_2, ..., A_n\}$, uses A to divide S into n subsets $\{S_1, S_2, ..., S_n\}$, and S_j is the sample that has A_j in A, S_{ij} is the sample number of class B_i in subset S_j. Info(S, A) is the needed information entropy. The calculation progress is as follows:

$$Info(S,A) = \sum_{j=1}^{n} \frac{S_{1j} + S_{2j} + ... + S_{nj}}{S} Info(A) \quad (5)$$

The split information SplitInfo (A) is the entropy of each value of attribute A about S, it is used to eliminate deviation of attribute that has a large number of value attribute. The calculation progress is

$$SplitInfo(A) = -\sum_{i=1}^{n} \frac{|S_j|}{|S|} \log\left(\frac{|SS_j|}{|S|}\right) \quad (6)$$

$$Gain(A) = Info(S) - Info(S,A) \quad (7)$$

$$GainRatio(A) = Gain(A) / SplitInfo(A) \quad (8)$$

Further, we also define indicator function as

$$\theta^t(i) = \begin{cases} 1, \text{sample } i \text{ is misclassified} \\ 0, \text{sample } i \text{ is classified rightly} \end{cases} \quad (9)$$

C. Random Forest Algorithm

Given a training set $X = x_1, x_2, ..., x_n$ with responses $Y = y_i$ through y_m, bagging repeatedly selects the training set and fits trees to these samples: For b =1 through B:

1. Sample, with replacement, n training examples from X, Y; call these $X_{b,} Y_b$
2. Train a decision or regression tree f_b on $X_{b,} Y_b$

When complete the training phase, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x') \quad (10)$$

Using cross-validation to find the optimal number of trees B, or by observing the out-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample [18].

III. PERFORMANCE EVALUATION MEASURES

In order to measures enables appropriate evaluation of

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 10, April 2014

proposed techniques. The performance of the proposed algorithms can be determined by the computation of total classification of accuracy, sensitivity, specificity, confusion matrix and the evaluation methods are defined as follow [19, 20].

$$\text{Classification Accuracy} = \left[\frac{\text{Correct classifed patterns}}{\text{Totalnumber of patterns}}\right] \quad (11)$$

$$Sensitivity = \left[\frac{TP}{TP + FN}(\%)\right] \quad (12)$$

$$Specificity = \left[\frac{TN}{FP + TN}(\%)\right] \quad (13)$$

The typical construction of the confusion matrix for the two classes is represented in Table 1. The difference between the actual patterns and the classified patterns is used to determine the performance of the proposed techniques [20].

Table 1. Representation of confusion matrix

	Classified Class	
Actual Class	1 (recurrent)	2 (non-recurrent)
1 (recurrent)	X_1	X_2
2 (non-recurrent)	Y_1	Y_2

In this study, the colorectal cancer dataset provided by the Chung Shan Medical University Hospital Tumor Registry was used in this study in order to verify the feasibility and effectiveness of C5.0, MARS and Random Forests (RF). Each patient in the dataset contains 25 predictor variables, namely, Sex, Primary Site, Histology, Behavior Code, Grade, Regional Lymph Nodes Examined, Regional Lymph Nodes Positive, Surgical Diagnostic and Staging Procedure at Other Facility, Surgical Diagnostic and Staging Procedure at This Facility, Clinical T, Clinical N, Clinical M, Clinical Stage Group, Pathologic T, Pathologic N, Pathologic M, Pathologic Stage Group, Surgical Margins of The Primary Site, Sequence of RT and Surgery, Sequence of Locoregional Therapy and Systemic Therapy, Dose of RT, Chemotherapy at Other Facility, Chemotherapy at This Facility, Vital Status, Cause of Death. And the response variable is recurrent or no. In this dataset of this study, there were totally 607 patients. Among them, 546 datasets with respect to the ratio of recurrent and non-recurrent patients were randomly selected as the training sample (estimating the parameters of the corresponding built classification models) while the remaining 61 will be retained as the testing sample (evaluating the classification capability of the built models). In the modeling of C5.0 classification model, the predictor variables should first be selected. Two significant independent variables were included in the final C5.0 model, namely Pathologic M, and Pathologic Stage Group. The classification results (the confusion matrix) of the testing sample using the obtained C5.0 model can be summarized in Table 2. From the results revealed in Table 2, we can observe that the average correct classification rate is 93.44% with 2 (2) class 1 (2) patients misclassified as class 2 (1) patients (Here a class 1 patient is defined as a patient with recurrent while a class 2 patient is a patient with non-recurrent).

Table 2. Classification results using C5.0

	Classified Class	
Actual Class	1 (recurrent)	2 (non-recurrent)
1 (recurrent)	27 (93.10%)	2 (6.90%)
2 (non-recurrent)	2 (6.25%)	30 (93.75%)
Average correct classification rate: 93.44%		

Table 3 shows the classification results of the testing sample using the obtained MARS model. The average correct classification rate is 85.25% with 5 (4) class 1 (2) patients misclassified as class 2 (1) patients.

Table 3. Classification results using MARS

	Classified Class	
Actual Class	1 (recurrent)	2 (non-recurrent)
1 (recurrent)	13 (72.22%)	5 (27.78%)
2 (non-recurrent)	4 (9.30%)	39 (90.70%)
Average correct classification rate: 85.25%		

In the modeling of the RF approach, all of the dataset are used for random feature selection and bootstrap data sampling. The result of testing dataset was decided by voting of the classification trees. From the result of Table 4 is observed the average correct classification rate is 78.69% with 8 (5) class 1 (2) patients misclassified as class 2 (1) patients.

Table 4 Classification results using RF

	Classified Class	
Actual Class	1 (recurrent)	2 (non-recurrent)
1 (recurrent)	25 (75.76%)	8 (24.24%)
2 (non-recurrent)	5 (17.86%)	23 (82.14%)
Average correct classification rate: 78.69%		

From Tables 2-4, it can be found that the average correct classification rates of the C5.0, MARS and RF models were 93.44%, 85.25% and 78.69%, respectively. The C5.0 model has the best classification capability in terms of the average correct classification rate. It outperforms the C5.0, MARS and RF models and hence provides an efficient alternative in conducting colorectal cancer classification tasks. In order to assess the robustness of these approaches, the performance of the C5.0, MARS and RF models was tested using 10 independent runs. Based on the findings in Table 5, the highest average correct classification rate for the Overall is 86.83% which is provided by the C5.0. Consequently, based on the results from this dataset, we can conclude that the C5.0 model is an effective alternative for colorectal cancer classification. In the table 5, after 10 runs, the selected important independent variables are Surgical Margins of The Primary Site, Pathologic Stage Group, Surgical Diagnostic and Staging Procedure at This Facility, and Pathologic M.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 10, April 2014

Table 5. Robustness evaluation of the MARS, C5.0, and RF

Approaches	Classification error %	Classification accuracy %	Sensitivity	Specificity
MARS	16.30	83.70	68.08	92.40
C5.0	13.17	86.83	70.42	94.61
RF	22.26	77.74	83.27	76.28

IV. CONCLUSION

As a result, our findings support that Surgical Margins of The Primary Site and Pathologic Stage Group are important and independent prognostic factor. In particular, Surgical Diagnostic and Staging Procedure at This Facility and Pathologic M were significantly related to the recurrence. Further, Surgical Margins of The Primary Site deeply invasive tumors and Pathologic Stage Group were independent risk factors. The presented results suggest the C5.0 decision tree is a good decision model. This is the first study using ensemble learning approaches for the analysis of risk factors for colorectal cancer, and results of this study will contribute to developing the clinical practice guideline for colorectal cancer.

REFERENCES

[1] Ong, L. S., Shepherd, B., Tong, L. C., Seow-Choen, F., Ho, Y. H., Tang, C. L., and Tan, K. The colorectal cancer recurrence support (CARES) system. Artificial Intelligence in Medicine, 11(3), 175-188, 1997.

[2] Leung, W. H., and Liu, C. K. Chemotherapy and Targeted Therapy in Colorectal Cancer: The Current Status. J. Cancer Res. Pract, 30(1), 11-20, 2014.

[3] Malcolm, A.W. Perencevich, N.P. Olson, R.M. Analysis of recurrence patterns following curative resection for carcinoma of the colon and rectum, Surg. Gynecol. Obstet. 152, 131-137, 1981.

[4] Hveem, T. S., Merok, M. A., Pretorius, M. E., Novelli, M., Bævre, M. S., Sjo, O. H., and Danielsen, H. E. (2014). Prognostic impact of genomic instability in colorectal cancer. British journal of cancer, 2014.

[5] Lengauer, C., Kinzler, K. W., and Vogelstein, B. Genetic instabilities in human cancers. Nature 396(6712): 643–649, 1998.

[6] Mouradov, D., Domingo, E., Gibbs, P., Jorissen, R. N., Li, S., Soo, P. Y., Lipton, L., Desai, J., Danielsen, H. E., Oukrif, D., Novelli, M., Yau, C., Holmes, C. C., Jones, I. T., McLaughlin, S., Molloy, P., Hawkins, N. J., Ward, R., Midgely, R., Kerr, D., Tomlinson, I. P., and Sieber, O. M. Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. Am J Gastroenterol 108(11), 1785–1793, 2013.

[7] Walker, A. S., Johnson, E. K., Maykel, J. A., Stojadinovic, A., Nissan, A., Brucher, B., and Steele, S. R. Future Directions for the Early Detection of Colorectal Cancer Recurrence. J Cancer, 5(4), 272-280, 2014.

[8] Cortet, M., Grimault, A., Cheynel, N., Lepage, C., Bouvier, A., and Faivre, J. Patterns of recurrence of obstructing colon cancers after surgery for cure: a population-based study. Colorectal Dis. May 2, 157-162, 2013.

[9] Paty, P. B., Nash, G. M., Baron, P. Long-term results of local excision for rectal cancer. Ann Surg. 236(4):522-529; discussion 529-530, 2012.

[10] Zhou, Z. H. Ensemble learning. In Encyclopedia of Biometrics 270-273, Springer US, 2009.

[11] Huang, X., Wang, H. N., and Li, L. P. Ensemble prediction model of solar proton events associated with solar flares and coronal mass ejections. Research in Astronomy and Astrophysics, 12(3), 313, 2012.

[12] Kruppa, J., Ziegler, A., Ko¨nig, I. R. Risk estimation and risk prediction using machine-learning methods. Hum Genet 131(10), 1639–1654, 2012.

[13] Hu Y. H., Wu F., Lo C. L., Tai, C. T. Predicting warfarin dosage from clinical data: a supervised learning approach. Artif Intell Med 56(1), 27–34, 2012.

[14] Multivariate Adaptive Regression Splines (MARS) https://www.statsoft.com/Textbook/Multivariate-Adaptive-Regression-Splines/button/2

[15] Gomathi, M., Thangaraj, P. A computer aided diagnosis system for lung cancer detection using machine learning technique. Eur J Sci Res 51:260–275, 2011.

[16] Malar, E., Kandaswamy, A., Chakravarthy, D., and Giri Dharan, A. A novel approach for detection and classification of mammographic micro calcifications using wavelet analysis and extreme learning machine. Comput Biol Med 42:898–905, 2012.

[17] Quinlan JR C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufinann, 1993.

[18] The machine learning technique: Random Forest http://en.wikipedia.org/wiki/Random_forest

[19] Watkins, A. B. Exploiting immunological metaphors in the development of serial, parallel, and distributed learning algorithms. PhD dissertation, University of Kent, Canterbury, March, 2005.

[20] VermaS, B., and Hassan, S. Z. Hybrid ensemble approach for classification. Applied Intelligence, 34(2), 258-278, 2011.