Cervical Cancer Prognosis using MARS and Classification

Farhana Kausar[1], Chetan J[2]

[1]Assistance Professor, Atria Institute of Technology, Visvesvaraya Technological University, Bangalore
[2]Student, Atria Institute of Technology, Visvesvaraya Technological University, Bangalore

---

***Abstract*—**This study applied advanced data mining techniques for recurrent cervical cancer in survival analysis. The medical records and pathology were accessible by the Chung Shan Medical University Hospital Tumor Registry. Following a literature review, expert consultation, and collection of patients' data, twelve variables studied included age, cell type, tumor grade, tumor size, pT, pStage, surgical margin involvement, LNM, Number of Fractions of Other RT, RT target Summary, Sequence of Locoregional Therapy and Systemic Therapy, LVSI. Two data mining approaches were considered where individuals are expected to experience repeated events, along with concomitant variables. After correcting for the four most important prognostic factors: pStage, Pathologic T, cell type and RT target Summary. Finally, clinical trials should randomize patients stratified by these prognostic factors, and precise assessment of recurrent status could improve outcome.

***Index Terms*—Recurrent event, cervical cancer, data mining technique.**

I. INTRODUCTION

One of the emerging issues in medical research is data mining. The widespread use of computers makes it easy to gather and manage large amounts of data from many different sources. A well-organized system can make available clinical, biological, genetic data, and all other information collected about patients. This data is often complex, meaning that it contains many elements related in non-obvious ways or characterized by explicit or implicit relationships and structures. Such integration is increasingly considered necessary in order to produce more accurate diagnoses. Cervical cancer remains one of the leading causes of cancer-related death among women globally [1], [2]. Even though the morbidity and the mortality have been decreasing in recent years, the morbidity rates of cervical cancer are the second leading type in women and the mortality rates are the sixth of the top ten cancers in Taiwan.

The natural history of cervical cancer begins with a normal epithelium which progress through various stages of dysplasia - cervical intraepithelial neoplasia grade CIN 1, CIN 2, CIN 3 - and finally, to invasive cervical cancer (ICC). There is a long time interval for the progression to ICC, and consensus on the fact that regression occurs in CIN. The most important part of therapy is to detect and eradicate local CIN 3 lesions before the progression to ICC and metastasis can occur. In general, the Papanicolaou (Pap) smear has been used widely as the most effective screening tool for detecting precancerous cervical lesions. Though screening and treat in its early phase, cervical cancer will be decreased significantly to its rate of incidence as well as death. Because cervical cancer is a cancer which can be controlled and avoided, the studies related to the causes of and the treatment to the cervical cancer has been described sufficiently in lots of advanced researches. On the other hand, there are few researches on its relationship between recurrent events and the mortality and incidence rate. Indeed, recurrent cervical cancer is a devastating disease for those women unfortunate enough to suffer such an event. Patients with recurrent disease or pelvic metastases have a poor prognosis with a 1-year survival rate between 15 and 20% [3].

When the recurrence is not surgically resectable, and/or suitable for curative radiation, therapeutic options are limited. In some advanced countries, the combination of cisplatin and topotecan is preferred since this is the only regimen which was able to show a statistical significant improvement of overall survival (OS) (9.4 months) without impairing quality of life due to intolerable toxicity [4]. But one has to be careful, because due to a change in primary therapy since 1999, when concomitant chemotherapy and radiotherapy became standard [4], [5], and due to the current investigation of the role of neo-adjuvant chemotherapy (EORTC 55994 (Cochrane Database of Systematic Reviews, 2004)), most people with recurrent cervical cancer will have had some challenge with a chemotherapeutic agent. This will influence responses in secondary treatment lines and will limit comparison of new studies with older ones including more chemonaive patients. Therefore, in the absence of surgical/ radiotherapeutic indications, chemotherapy should be targeted to the prolongation of survival with minimum morbidity and to the improvement of subjective symptoms, thus preserving quality of life. Unfortunately, in these conditions, there is no evidence of a significant impact on survival or on quality of life. For these reasons, the role of chemotherapy in recurrent disease remains to be defined and the search for more active and less toxic agents must be continued. Since, the treatment of recurrent cervical carcinoma is still a clinical challenge.

II. METHODS

In generally, medical data analyses have been performed employing standard statistical methods, since clinicians can usually better understand them and are often already familiar with some of the statistical packages that are widely available. Despite their popularity, however, many statistical techniques are based on very simple models, which often fail

to catch data complexity. In this light, data mining can provide quite useful tools, since its models are usually much more powerful and flexible and can actually tackle problems with complex data. In the health field, data mining applications have been growing considerably as it can be used to directly derive patterns, which are relevant to forecast different risk groups among the patients. To the best of our knowledge data mining technique such as classification has not been used to analyse the recurrence cervical cancer. Hence, in this paper we made an attempt to identify patterns from the database of the cervical cancer patients using several advances techniques as follows.

A. MARS

Multivariate Adaptive Regression Splines (MARS) was first introduced by Friedman [6] to efficiently approximate the relationship between a dependent variable and a set of explanatory variables in a piece-wise regression. Capability of MARS for modeling time series data was subsequently demonstrated in [7], where lagged values of the time series were treated as explanatory variables. In recent years, application of MARS has been reported for modeling a variety of data, such as, speech modeling [8], mobile radio channels prediction [9], and intrusion detection in information systems security [10].

In addition, MARS was employed to model the relationship between retention indices and molecular descriptors of alkanes [10], and to describe pesticide transport in soils [11]. MARS has also been applied to predict the average monthly foreign exchange rates [12], to model credit scoring [13], and for data mining on breast cancer pattern [14]. In all of the cited studies, promising results have been reported where MARS has been employed either for forecasting or for data mining purposes.

MARS is a non-parametric modeling approach versus the well-known global parametric modeling methods such as linear regression [6]. In global parametric approaches the underlying relationship between a target variable and a set of explanatory variables is approximated using a (usually simple) global parametric function which is fitted to the available data. While global parametric modeling methods are relatively easy to develop and interpret, they have a limited flexibility and work well only when the true underlying relationship is close to the pre-specified approximated function in the model. To overcome the weaknesses of global parametric approaches, non-parametric models are developed locally over specific subregions of the data; the data is searched for optimum number of subregions and a simple function is optimally fit to the realizations in each subregion.

B. C5.0

C5.0 classifier is a process for the classification and analysis of information hidden in large datasets/databases, which retrieves useful information in the form of a decision tree, i.e., a flowchart like tree structure [15]. The algorithm adopts 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 more noise [15]. C4.5 [16] builds decision trees from a set of training data in the same way as ID3 (Iterative Dichotomiser 3), using the concept of information entropy. The training data is a set of already classified samples. Each sample is a vector including attributes or features. The training data is augmented with a vector representing the class that each sample belongs to. Each attribute of the data can be used to make a decision. C4.5 examines the normalized information gain that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller sub-lists.

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 as follows:

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

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:

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

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 as follows:

$$SplitInfo(A) = -\sum_{i=1}^{n} \frac{|S_j|}{|S|} \log\left(\frac{|S_j|}{|S|}\right) \quad (3)$$

$$Gain(A) = Info(S) - Info(S,A) \quad (4)$$

$$GainRatio(A) = Gain(A) / SplitInfo(A) \quad (5)$$

III. EMPIRICAL STUDY

In this study, the cervical cancer dataset provided by the Chung Shan Medical University Hospital Tumor Registry is used in this study order to verify the feasibility and effectiveness of C5.0 and MARS. Each patient in the dataset contains 12 predictor variables, namely, age, cell type, tumor grade, tumor size, pT, pStage, surgical margin involvement, lymph node metastases(LNM), Number of Fractions of Other RT, RT target Summary, Sequence of Locoregional Therapy and Systemic Therapy, lympho-vascular space

involvement(LVSI). And the response variable is recurrent or no). There are totally 168 patients in the dataset. Among them, 118 datasets with respect to the ratio of recurrent and non-recurrent patients (the prior probabilities or simply priors) were randomly selected as the training sample (estimating the parameters of the corresponding built classification models) while the remaining 50 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 (or independent) variables should first be selected. Two significant independent variables were included in the final C5.0 model, namely pT and RT target Summary. The classification results: the average correct classification rate is 96.0%. For modeling the MARS classification model, all of the twelve predictor variables are used as inputs. The classification results: the average correct classification rate is 86.00%.

From Table I, it can be found that the average correct classification rates of the C5.0 and MARS models are 96.00%, and 86.00%. The C5.0 model has the best classification capability in terms of the average correct classification rate. It outperforms the SVM model and hence provides an efficient alternative in conducting cervical cancer classification tasks.

In order to assess the robustness of the C5.0 method, the performance of the C5.0 and MARS model was tested using 10 independent runs. Based on the findings, the C5.0 model not only generates the better classification result, but also can be used to select important independent variables for cervical cancer classification. The selected important independent variables can provide useful information for cervical cancer treatment. In this study, after 10 runs, the selected important independent variables are pStage, pT, cell type, and RT target Summary.

TABLE I: ROBUSTNESS EVALUATION OF THE C5.0 AND MARS

	{1-1}		{2-2}		Overall	
Model Runs	C5.0	MARS	C5.0	MARS	C5.0	MARS
1	100.00	88.24	87.50	81.25	96.00	86.00
2	91.67	97.22	100.00	85.71	94.00	94.00
3	95.00	90.00	80.00	70.00	96.00	86.00
4	89.47	89.47	100.00	75.00	92.00	86.00
5	91.89	97.30	92.31	76.92	92.00	92.00
6	94.87	87.18	81.82	90.91	92.00	88.00
7	84.38	96.88	94.44	33.33	88.00	74.00
8	97.14	94.29	80.00	53.33	92.00	82.00
9	95.12	95.12	100.00	77.78	96.00	92.00
10	88.89	88.89	92.86	78.57	90.00	86.00
Average	92.05	92.46	91.27	72.28	92.44	86.60

IV. DISCUSSION

Cervical cancer is a type of cancer that starts in the cervix, the lower part of the uterus that opens at the top of the vagina. Most of the time, early cervical cancer has no symptoms. In this research our aim is to identify the significant risk factors for the recurrent cervical cancer. These include tumor size, lymphovascular space involvement, depth of tumor invasion, lymph node metastasis, and parametrial involvement. Prognostic models combining these factors were also devised [17], [18]. These factors are interrelated, but analyses using these factors do not reflect the true prognoses. However, these models were still insufficient to provide individualized risk assessments. In our study, pStage deeply invasive tumors, and pT were independent risk factors, in contrast to other similar analyses [19]. Our findings support that pStage and pT are important and independent prognostic factor. Cell type and RT target Summary were significantly related to the recurrence.

V. CONCLUSION

This study has demonstrated the use of data mining to identify some specific risk factors for the recurrent cervical cancer problem. The presented results suggest the decision tree is a good decision model. It is simple and preserves all the advantages of classical decision trees. For medical interpretation, however, cooperation with doctors is needed to verify the model build. Perhaps by analysis existing or easily measured data about a patient we can develop some results by which a physician caring a patient can better decide when to take the treatment. Further, the goal is to afford the patient the opportunity to have a reasonable quality of life in addition to providing the chance for a cure in the future.

REFERENCES

[1] D. M. Parkin, F. I. Bray, and S. S. Devesa, "Cancer burden in the year 2000: the global picture," *Eur J Cancer*, vol. 37 (suppl), pp. S4-S66, Oct. 2001.

[2] S. J. Goldie, L. Kuhn, L. Denny, A. Pollack, and T. Wright, "Policy analysis of cervical cancer screening strategies in low-resource setting: clinical benefits and cost effectiveness," *JAMA.* vol. 285, pp. 3107-3115, Feb. 2001.

[3] J. S. Berek and N. F. Hacker, *Practical Gynaecologic Oncology*, New York: Lippincott Williams & Wilkins, 2005.

[4] S. E. Waggoner, "Cervical cancer," *Lancet*, vol. 361, pp. 2217-2225, 2003.

[5] C. H. Lai, J. H. Hong, and S . Hsueh, "Preoperative prognostic variables and the impact of postoperative adjuvant therapy on the outcomes of stage IB or II cervical carcinoma patients with or without pelvic lymph node metastases," *Cancer*, vol. 85, pp. 1537-1546, 1999.

[6] J. H. Friedman, "Multivariate adaptive regression splines," *The Annual of Statistics*, vol. 19, pp. 1-67, 1991.

[7] P. A. W. Lewis and J. G. Stevens, "Nonlinear modeling of time series using multivariate adaptive regression splines," *Journal of American Statistical Association*, vol. 86, no. 416, pp. 864-877, 1991.

[8] H. Haas and G. Kubin, "A multi-band nonlinear oscillator model for speech," in *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems & Computers*, vol. 1, pp. 338-342, 1998.

[9] T. Ekman and G. Kubin, "Nonlinear prediction of mobile radio channels: measurements and mars model designs," in *IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing*, Arizona, 1999, vol. 5, pp. 2667-2670.

[10] Q. S. Xu, D. Massart, Y. Z. Liang, and K.-T. Fang, "Two-step multivariate adaptive regression splines for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors," *Journal of Chromatography*, vol. 998, no. 1-2, pp. 155-167, 2003.

[11] C. C. Yang, S. O. Prasher, R. Lacroix, and S. H. Kim, "A multivariate adaptive regression splines model for simulation of pesticide transport in soils," *Biosystems Engineering*, vol. 86, no. 1, pp. 9-15, 2003.

[12] A. Abraham, "Analysis of hybrid soft and hard computing techniques for forex monitoring systems," in *IEEE Proc. International Conference on Fuzzy Systems*, Honolulu, 2002, vol. 2, pp. 1616-1622.

[13] T. S. Lee and I. F. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines," *Expert Systems with Application*s, vol. 28, pp. 743-752, 2005.

[14] S. M. Chou, T. S. Lee, Y. E. Shao, and I.-F. Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," *Expert Systems with Applications*, vol. 27, pp. 133-142, 2004.

[15] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, 2005.

[16] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufinann, 1993.
[17] R. E. Estape, R. Angioli, M. Madrigal, M. Janicek, C. Gomez, and M. Penalver, "Close vaginal margins as a prognostic factor after radical hysterectomy," *Gynecol Oncol*, vol. 68, pp. 229-32, 1998.
[18] G. Delgado, B. Bundy, R. Zaino, B. U. Sevin, and W. T. Creasman, "Major Prospective surgical – pathological study of disease-free interval in patients with stage Ib squamous cell carcinoma of the cervix: a gynecologic oncology group study," *Gynecol Oncol*, vol. 38, pp. 352-357, 1990.
[19] T. Kamura, N. Tsukamoto, N. Tsuruchi, T. Saito, T. Matsuyama, and K. Akazawa, "Multivariate analysis of the histopathologic prognostic factors of cervical cancer in patients undergoing radical hysterectomy, " *Cancer*, vol. 69, pp. 181-186, 1992.