Integrated Intelligent Research (IIR)

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015 Page No.63-66
ISSN: 2278-2419

A Pioneering Cervical Cancer Prediction Prototype in Medical Data Mining using Clustering Pattern

R.Vidya[1],G.M.Nasira[2]
[1]Research scholar of M.S.University,Tirunelveli, Assistant Professor /Department of Computer Science
St.Joseph's college, Cuddalore, India
[2]Assistant Professor/Department of Computer Science, ChikkannaGovernment College for Women, Tirupur, India
Email: vidya.sjc@gmail.com, nasiragm99@yahoo.com

Abstract—Let us not make the cure of the disease more unbearable than the disease itself this quote is the most durable and inspirational line of medicine field. Data mining is said to be an umbrella term which refers to the progression of finding out the patterns in data. This can be even succeeded typically with an assistance of authoritative algorithm to automate search (as a part). This paper reveals out, how the C2P (Cervical Cancer Prediction) model is approached by a data mining algorithm for prediction. The prediction of C2 (Cervical Cancer) has been a challenging problem in research field. In the Data mining applications, we are utilizing RFT (Random Forest Tree) algorithm to do the prediction. To the best of our knowledge, we use popular clustering K-means technique to achieve more accuracy.

Keywords— Cervical Cancer, Random Forest Tree, K-means learning, Data mining, Clustering K-means.

I. INTRODUCTION

Cancer of the cervix is one of the most prevalent forms of cancer worldwide, the major burden of the disease being felt in developing countries like India. Cervical carcinoma still continues to be the most common cancer among women and accounts for the maximum cancer deaths each year. Persistent infections with High-Risk (HR) Human Papilloma viruses, such as HPV 16,18,31,33 and 45 have been identified as a major development of the distance [8]. This model interpretability and prediction accuracy provided by Random Forest is very unique among popular machine learning methods. Accurate predictions and better generalizations are achieved due to utilization of ensemble strategies and random sampling.Random Forest three main features that gained focus are: (i) Accuratepredictions results for a variety of applications. (ii) Through model training, the importance of each feature can be measured (iii) Trained model can measure the pair-wise proximity between the samples. The research of our paper is as followed by cervical cancer and diseases and then proceeded by medical data mining and RFT where we deal with K-Mean algorithm in the next phase. Followed by this we have given out our C^2P Model along with how to prevent cancer. Later the paper concludes by Result and conclusion. Characterized by abnormal bleeding, pelvic pain and unusual heavy discharge, the disease develops in the tissues of the cervix, a part connecting the upper body of the uterus to the vagina.

II. CERVICAL CANCER DISEASE AND ITS SYMPTOMS

In worldwide C^2 is the third most common cancer in women and seventh overall. Majority of this global burden is felt in low and middle income developing countries (WHO, 2010) and in low socio-economic groups within countries [1].It comprises of endocervix or the upper part which is close to the uterus and covered by glandular cells; and the ecocervix, the lower part which is close to the vagina and covered by Seamus cells. The two regions of the cervix meet at the "transformation zone". It is this region where most cervical cancer begins to develop [6].

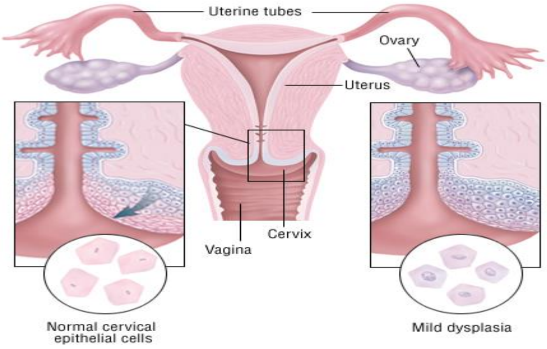


Figure 1. Cervix Cancer with its Cells (Developing Stage)

A. Signs & Symptoms of Cervical Cancer

Early cervical cancer has no symptoms and it cannot be identified in a clear way. Symptoms do not begin until the pre-cancer grows to an aggressive stage and starts to spread to nearby tissue. When this happens the most common symptoms are:

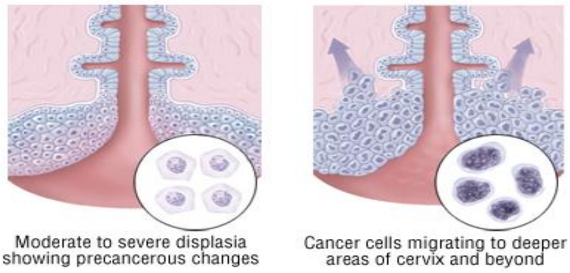

Moderate to severe displasia showing precancerous changes
Cancer cells migrating to deeper areas of cervix and beyond

Figure 2. Cancer Cells in Progressive Stage

- Vaginal bleeding abnormally primarily bleeding after sex, after menopause, spotting between periods, long or heavy periods than normal, bleeding after pelvic exam.
- Unusual discharge from vagina- the discharge may contain blood.
- Vaginal intercourse
- These symptoms can be caused other than cervical cancer too but anyway having a check on it will be good.

III. MEDICAL DATA MINING AND RFT

RFT is best among classification which is able to classify and predict large amount of data with accuracy[11]. Random Forest Tree gives an opportunity to set parameter there is no need to prune tree. Accuracy and variable will be generated automatically. The outliers are very sensitive in training data. Different subset of training dataset are selected (2/3) with replacement to train each tree. Remaining training data (OOB) are used to estimate error and variable importance class assignment is made by the number of votes from all of the trees and for regression the average of the result is used.

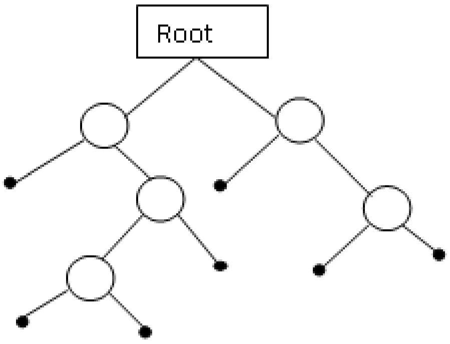


Figure 3. Structure of RFT

RFT algorithm was developed by Leo Breiman and Adelecutler. RFT grows into many classification trees.Each tree is grown as follow [10].

- If number of training set cases set are N, Sample N are collected at random but with the replacement from original data. This sample will be training tree
- M input variables a number MM is specified such that each node, m variable are selected at random out of M
- Best split on this m is used in split the node. The value of m is detained continuously during the forest growing
- Every tree is grown to the largest extent possible so there is no pruning

IV. CLUSTERING K-MEANS ALGORITHM

Whitening is a common preprocess in deep learning work which provides better performance in respect to the choice of unsupervised learning algorithm. Preprocessing is a common practice forperforming many simple normalization steps before attaining its final stage[4]. We apply a K-Mean clustering to learn K centroids *$C^{(k)}$*from the input data. Given the learned centroids*$C^{(k,}$*we consider two choices for the feature mapping *f*,

$$f_k(x\ ) = \begin{cases} 1 & \text{if } K = \arg\min;\ \|\ c^{(i)} - x\ \|^2 \\ 0 & \text{otherwise} \end{cases}$$

This is called K-means hard.The second is a non-linear mapping that attempts to be “softer” than the above encoding but also yield sparse outputs through simple arithmetic,

$$f_k(x) = \max\{0,\ u(z) - z_k\} \text{ where } z_k = \|x\text{-}c^{(k)}\|^2 \text{ and } u(2)$$

Is the mean of the elements of z. This is called K-means triangle[3].

K-Means Algorithm Unsupervised-Learning-The k-means algorithm is build upon the following operations

Step 1-Choose initial cluster centers C1, C2, …, C_K randomly from the n points.W1,W2,WN, WI € Rm.

Step 2-Assign point W_i, I=1,2,…,N to cluster Z_j= 1,2,…K if and only if $\|w_q - c_q\|$, P=1,2,…K and J ≠ P Ties are resolved arbitrarily.

Step 3-Compute *the new cluster centers*
$C_1^{*}, C_2^{*}, \ldots, C_k^{*}$ as following
$C_1^{*} = (I/n)\ \Sigma W_j \qquad I = 1,2,\ldots K$
$W_j \in ZJ$

Step 4-If $C_i^{*} = C_i$, I = 1,2,…K then terminate otherwise $C_iC_j^{*}$ and go to step 2.

Data mining provides the methodology and technology to analysis the useful information of data[3]. Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets. [7] K-mean is a greedy algorithm; can only converge to a local minimum clustering algorithm is divided into two categories. Partition and hierarchical clustering algorithm K-means partitioned one [2] Proposed by the queen in 1967. K-means is numerical unsupervised non-deterministic interactive method [5].

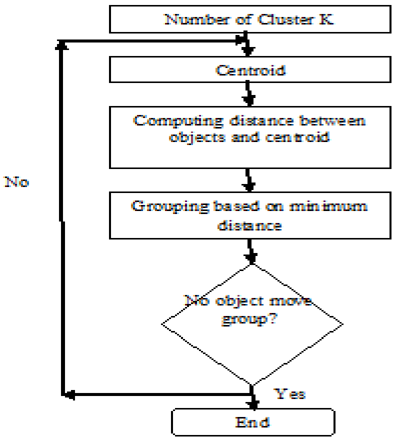

Number of Cluster K
Centroid
Computing distance between objects and centroid
Grouping based on minimum distance
No object move group?
No
Yes
End

Figure 4. Process of K-means Learning

This table 1 gives the result of Logistic analysis for regression. In total 11 risk factors are included, the main factors which gives up to fatal rise of problems like Illiteracy, Married Life, Early Menarche, Multiple Pair and Poor Hygiene. Using the linear transform the statistical weight for the risk factors is identified. The statistical weight analysis as followed with figure 6 of statistical weight of risk factor. Table 2 shows the classification of data subjects which thethe identification of cases and controls. The total risk category is used on a score of 7 with the interval limit. The classification based on control is as follow

V. C^{2}P MODEL – DATA SET AND RISK GROUP DETECTION

Risk factors can be classified in to four categories. Category I contains white discharge, hip-pain, smelly vaginal discharge, early marriage, sex at early age and multiple sexual partners. Category II includes risk factors likely malnutrition, male sexual behaviors, husband's food habit and living condition. Category III – (psychological factors) Category IV – (age, family history)[9].

Architecure of C^{2}P model

VI. PREVENTING CANCER COUNT ON

The cervical cancer does not have any symptoms until it reaches an aggressive stage. So a minor problem should be taken into account like

- Unbalanced vaginal Bleeding
- Vaginal Discharge with Odor
- Vaginal Discharge in a Watery form
- Vaginal Discharge along with blood

A. Preventive steps

- General education for women population about life style factors.
- Risk factor screening in general
- Collecting information like age, family history, menopause period & husband's activities.
- Collecting results of laboratory texts including Pap smear test and biopsy test.
- Collecting direct investigation reports like VILLA etc.
- Giving counsel to go the nearest hospital for the screening regularly C2 dataset is taken from open source data online. The data are then clustered with k-means algorithm using MATLAB.

TABLE I. RESULT OF NUMEROUS LOGISTIC REGRESSION ANALYSIS AND STATISTICAL WEIGHTINESS OF THE SIGNIFICANT RISK FACTOR

Risk Factor	Regression Co-Efficient	Odds Ratio	95% of RC for OR	Statistical Weight
Literacy	1.2345	3.527	2.207 -5.556	13
Married Life	0.3385	1.699	1.439 -3.494	7
Early Menarche	0.9795	1.402	0.889-2.212	5
Marital Status	0.3327	2.663	0.708-1.873	2
Genetal Infection	0.1410	2.145	0.453-1.190	5
Multiple Pair	0.0675	1.151	1.354-3.398	7
Abortion	0.0448	1.264	0.660-1.732	2
Tobacco Use	0.637	0.734	1.693-4.190	6
Poor hygiene	0.5299	1.499	0.708-1.873	10

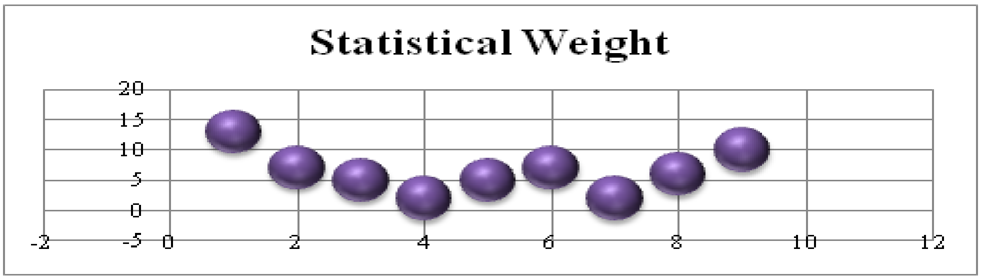


Figure 5. Statistical Weight of Risk Factor

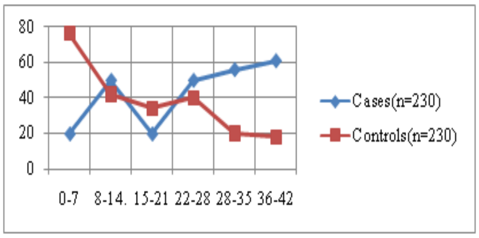


Figure 6. Score Control Analysis for Case

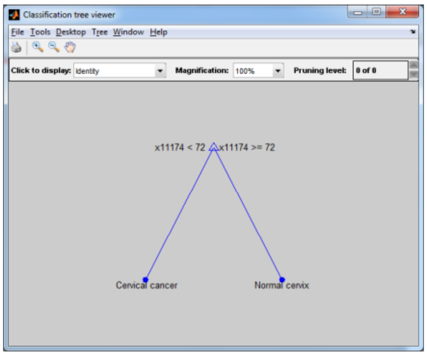
Score	Cases (n=230)	Controls (n=230)
0-7	20	76
8-14	50	42
15-21	20	34
22-28	50	40
28-35	56	20
36-42	61	18

VII. RESULT

With RFT- data mining algorithm we achieved 93.37% accuracy. But, with enhancement, i.e., with k-means learning we achieved 97.37% result in our research work.

TABLE II. DATA MINING ACCURACY

SL.NO	DATA-MINING ALGORITHM	ACCURACY
1	RFT	93.54
2	RFT WITH K-MEANS LEARNING	96.77


Classification tree viewer

Click to display: Identity
Magnification: 100%
Pruning level: 0 of 0

Figure 7. Result of accuracy using RFT

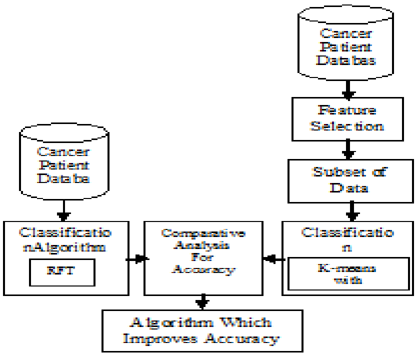


Figure 8. Architecure of C^2P model

VIII. CONCLUSION

Cervical cancer ranks as the first most frequent cancer among women in India. Tree predictors make a combination where every tree depends on the random vector sampled value independently with the same distribution for all the trees is said to be Random Forest. It is a perfect tool for building guesses considering overfit if there or not so to avoid large numbers. In this work, K-means learning was introduced to achieve better accuracy.Experimental results conducted on the collected cervical cancer data set demonstrated the effectiveness of the proposed techniques. Particularly, the proposed method mainly aims to predict the cancer cervix which is usually done by data mining algorithm. This paper is expected to be of benefit for medical decision making systems to give an alternative choice for medical practitioners to construct more accurate predictive models and stronger classifiers.We are bound to thank the Doctors of the JIPMER Hospital, Puducherry for their information. It was very useful and the information's provided by them was clear and made us to create a vivid picturization of our complications.

REFERENCES

[1] Rao, Y. N., Sudhir Gupta, and S. P. Agarwal. "National Cancer Control Programme: Current Status & Strategies." Fifty Years of Cancer Control In India. Dir Gen of Health Services, MOHFW, Government of India (2002): 41-7.
[2] Thangavel, Kuttiannan, P. Palanichamy Jaganathan, and P. O. Easmi. "Data mining approach to cervical cancer patients analysis using clustering technique." Asian Journal of Information Technology 5.4 (2006): 413-417.
[3] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." ACM SIGMOD Record. Vol. 22. No. 2. ACM, 1993.
[4] Selim, Shokri Z., and Mohamed A. Ismail. "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality." Pattern Analysis and Machine Intelligence, IEEE Transactions on 1 (1984): 81-87.
[5] Sun, Geng, et al. "Research on K-means Clustering Algorithm." Journal of Changchun Normal University 2 (2011): 001.
[6] Saslow, Debbie, et al. "American Cancer Society guideline for the early detection of cervical neoplasia and cancer." CA: a cancer journal for clinicians 52.6 (2002): 342-362.
[7] Ananthanarayana, V. S., M. Narasimha Murty, and D. K. Subramanian. "Efficient clustering of large data sets." Pattern Recognition 34.12 (2001): 2561-2563.
[8] Petignat, Patrick, and Michel Roy. "Diagnosis and management of cervical cancer." BMJ: British Medical Journal 335.7623 (2007): 765.
[9] American Cancer Society (accessed 9/5/07) some information taken from www.cancer.org
[10] Ali, Jehad, et al. "Random forests and decision trees." International Journal of Computer Science Issues (IJCSI) 9.5 (2012).
[11] Random Forest Algorithm