PREDICTING CERVICAL CANCER USING MULTIPLE MACHINE LEARNING TECHNIQUES

Dilip Kumar Baruah
Assistant Professor, Department of Computer Science, Moridhal College, Moridhal, Dhemaji Assam, India

Neelutpol Gogoi
Assistant Professor, Department of Computer Science, Moridhal College, Moridhal, Dhemaji Assam, India

ABSTRACT

Now a day, cervical cancer is the most common and prevailing gynecologic malignancies. Cervical cancer is the third type of cancer after breast and lungs cancer in women. Although its highly preventable disease provided early screening is done so as to minimize the global burden. However, due to unawareness, ignorance, lack of medical facilities and expensive procedures in developing countries, the vulnerable patient populations cannot afford to undergo examination regularly. A novel ensemble approach is presented in this paper to predict the risk of cervical cancer using machine learning approach. In this paper, we proposed a prediction model that can predict with accuracy the presence or absence of cervical cancer from as many as 35 possible risk features recorded for each woman. As per the results, the proposed approach is accurate, scalable and practical.

Keywords: Cervical cancer, Diagnosis, prediction, imbalance.


1. INTRODUCTION

Cancer is a generic term for a large group of diseases that can affect any part of the body and at times could take the life of patients. Cancer can be reduced if cases are timely detected and treated as early as possible [1]. Cervical cancer is a common type of cancer prevailing in women these days. Women especially in developing countries, there are so restricted resources. Additionally, sometimes patients do not take care to routine screening. Therefore, the most important problems during diagnosis are determination of the finest screening plan and
estimation of individual risks of each patient. In most of these screening methods results have been highly correlated with the experience of the physician and its subjective decision [2]

2. RELATED WORK

Machine learning has been of great help in many medical application and be used as a classifier in the early detection of the cancerous cell present in the cervix region of the uterus. To analyze medical dataset in machine learning algorithm, this study includes Intelligence data analysis [3]. There are various model and methodologies used in broadcast of cervical cancer by segmentation. Classification for cervical cancers and into the different categories, such as machine learning algorithm like Support Vector Machine (SVM), Grey Level Co Occurrence(GLCM) Matrix, Nearest Neighbours’ (KN-N) and a Multi Weighted Adaptive Regression Spines (MARS) Conventional Neural Network(CNN) spritul Fuzzy Clustering algorithm, probabilities neural network (PNS), Genetic Algorithm, Random Forest Tree (RFT), Classification and Regression tree, Clustering Algorithm for feature extraction segmentation and classification. This paper also include different data set related to the cervical cancer and present a review on the computational method [3].Segmenting a cell region containing either is single cell or multiple overlapping circles in to obtain a Corresponding segmentation for each nuclear and capture the true structure of that nuclear cell. The first phase constant of Segmentation and images by a nonparametric helical segmentation algorithm that is spiritual and shape information as well the gradient information and the second phase obtain nucleus region and brighter paints area by the classification of the segmenting resulting from first phase based on their spiritual and shape features [4][5].

3. ARCHITECTURE OF METHODLOGY

4. METHODOLOGY

Step 1. Collecting medical data set.

Step 2. Preprocessing of medical data set, removing noisy data.

Step 3. Extract feature from dataset.

Step 4. Signal prediction, which will be used in our Machine learning algorithms.
Step 5. Generate graph.

Step 6. Evaluate the accuracy of our algorithms.

4. DATASET

Machine learning requires maximum value of data which must be structural data. But in cervical cancer several patients decided not to answer some of the questions because of privacy concerns (missing values). The dataset was collected from Hospital Universitario de Caracas in Caracas Venezuela. The dataset comprise 35 features (table 1) that include demographic information, habits and historic medical records 858 patients. Our approach is to follow the OSEMN protocol in which, (Obtain) the dataset will be obtained from the University of California at Irvine (UCI) machine learning repository, (Scrub) after an initial inspection of the dataset to identify missing values and any other obvious issues we will develop and implement processes to minimize the negative impacts of those data values. The asymptomatic nature is the major challenge faced in the diagnosis of cervical at early stage. (Explore) EDA (Exploratory Data Analysis) will be worked out to understand the data its range of the values within each feature and any Correlations between features to help predict the target.

Table 1. Attribute Information

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Feature Information</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Int</td>
<td>STD: pelvic inflammatory Disease</td>
<td>Bool</td>
</tr>
<tr>
<td># of partners</td>
<td>Int</td>
<td>STD: genital herpes</td>
<td>Bool</td>
</tr>
<tr>
<td>Age of 1st intercourse</td>
<td>Int</td>
<td>STD: mollusccontagiorum</td>
<td>Bool</td>
</tr>
<tr>
<td># of pregnancies</td>
<td>Int</td>
<td>STD: AIDS</td>
<td>Bool</td>
</tr>
<tr>
<td>Smokes</td>
<td>Bool</td>
<td>STD: HIV</td>
<td>Bool</td>
</tr>
<tr>
<td>Smokes year</td>
<td>Int</td>
<td>STDs: Hepatitis</td>
<td>Bool</td>
</tr>
<tr>
<td>Smokes pack/year</td>
<td>Int</td>
<td>STD: HPV</td>
<td>Bool</td>
</tr>
<tr>
<td>Hormonal Contraceptives</td>
<td>Bool</td>
<td>STD: Number of Diagnosis</td>
<td>Int</td>
</tr>
<tr>
<td>Hormonal Contraceptives Years</td>
<td>Int</td>
<td>STD: Time Since first diagnosis</td>
<td>Int</td>
</tr>
<tr>
<td>IUD</td>
<td>Bool</td>
<td>STD: Time since last diagnosis</td>
<td>Int</td>
</tr>
</tbody>
</table>

5. PRE-PROCESSING

The following features have integer values and all have missing values in 1 to 12% of records:

- Age
- Number of partners
- Age of 1st intercourse
- Number of pregnancies
- Smokes (years), Smokes(packs/years)
- Hormonal Contraceptives (years)
6. FEATURE EXTRACTION

Number of sexual partners - multiple sexual partners is one of the risk factors of human infection which is a major cause of cervical cancer. It is not clear whether the number of sexual partners is an independent risk factor for cervical cancer.

This graph represents mostly no of sexual partners are two. The risk of both content malignant and non-malignant diseases is stable in women with more than 3 to 7 sexual partners.

6.1. First Sexual Intercourse

Early age sexual intercourse may be a big reason of cervical cancer in women. The graph clearly shows that the girls/women having intercourse at an early stage of their lives are at great risk of cervical cancer.

6.2. Hormonal Contraceptives (years)

Continuous intake of hormonal contraceptives for certain years from around age 20 to 30 years is estimated to increase the cumulative incidence of invasive cervical cancer. The graph for the same is given in below. Fig.3.
6.3. IUD

Women who use intrauterine devices (IUDs) for birth control, even for a short time, have a lower cervical cancer risk. The graph for the same is given in Fig. 4.

7. RESULT AND DISCUSSION

Figure 4 Hormonal Contraceptives (years)

Figure 5 IUD years

Figure 6 Correlation of 12 features
In this study, the models were evaluated based on performance measures accuracy of two models, Logistic regression and Random forest. The experimental were done with training data obtained from a web source. The same training and testing dataset is utilized in experimental model. It is a machine learning model; the understanding is easy to compression of other model. We achieve excellent results comparing two algorithms. The accuracy in our model is 99.38%, total runtime for grid search on random forest classifier is 12.9 second and a recall score 0.875 was reached in both the models.

Schiller, Hinselmann and cytology are features representing medical diagnostic tools used to identify cervical cancer.

8. CONCLUSION
This paper presents the compression study in between different machine learning algorithms. Different pre-processing technique has been used to address the imbalance data. Logistic Regression and Random forest combine both techniques. We found that random forest technique is better result than other one due to higher accuracy obtained by our model. We obtain the best overall result with accuracy 99.38% by selecting features such as age, number of sexual partners, first sexual intercourse, number of pregnancies, smokes, Hormonal Contraceptives, IUD, STD (number), STD’s (number of diagnosis) STD’s (Time since first diagnosis), STD’s (Time since last diagnosis) recall score of both model is 0.875. We prefer python classifiers. In the future work we would like to try more algorithms for better results.

REFERENCES

AUTHORS PROFILE

Dilip Kumar Baruah, MCA, Assistant Professor, Department of Computer Science, Moridhal College, Moridhal, Dhemaji Assam, India

Neelutpol Gogoi, MTech, Assistant Professor, Department of Computer Science, Moridhal College, Moridhal, Dhemaji Assam, India