Evaluation of Recursive Feature Elimination and LASSO Regularization based optimized feature selection approaches for cervical cancer prediction.

Authors:
Mohamed Hamada,
Presenter: Jesse Jeremiah Tanimu,
Mohammed Hassan,
Habeebah Kakudi,
Patience Robert

November, 2021
Overview

1. Introduction
   - Machine Learning concepts & Techniques
   - Cervical Cancer
2. Problem Statement
3. Aim and objectives
4. Related Work
5. Methodology
6. Evaluation Metrics
7. Results
8. Research Architecture
9. Research contribution
10. Summary and conclusion
11. Future Work
12. References
Introduction

- **Machine Learning concepts and techniques**
- Machine learning is a branch of AI. It is a technology that enables systems to learn directly from examples, data, and experience. The system might learn and improve with experience, and with time.
Cervical cancer

- Cervical cancer is a malignant tumor of the cervix. The lowermost part of the uterus.
- Every year more than 270,000 women die from cervical cancer, more than 85% of these deaths are in developing countries (WHO).
- With estimated new cases of 444,500 annually (IARC).
- Ranked as the second most frequent malignant tumor among women in Nigeria.
- When detected in early stages, it can often be cured by removing the afflicted tissues.
Problem Statement

- There have been many attempts to use machine learning (ML) in tasks such as prediction to help in the early detection of cervical cancer.
- However, these models while being able to predict outcomes of cervical cancer still suffer from one or more of the following limitations: low application of dimensionality reduction techniques, and solving the problem of overfitting in decision tree (DT).
- Hence new approach to develop a machine learning model that will select the most relevant features to detect cervical cancer at an early stage can offer the potential to tackle cervical cancer in a more cyclopedic approach and build a healthier future for girls and women.
Aim and Objectives

AIM

- The paper focuses mainly on predicting the biopsy results of patients with cervical cancer. It concentrates majorly in selecting the subset of features that will aid in the early prediction of cervical cancer.

OBJECTIVES

- Two feature selection techniques which include recursive feature elimination (RFE), and least absolute shrinkage and selection operator (LASSO) have been proposed to identify the most relevant subset of features in the prediction of cervical cancer.

- Finally, to compare the results of the 2 feature selection techniques.
## Related work

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Techniques</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatlawi (2017)</td>
<td>DT</td>
<td>The model was able to classify and identify patients with cervical cancer.</td>
<td>No feature selection Technique applied.</td>
</tr>
<tr>
<td>Wu and Zhou (2017)</td>
<td>SVM Variants</td>
<td>SVM and its variants have shown to be extremely powerful classifiers.</td>
<td>SVM and its variants can be slow to train.</td>
</tr>
<tr>
<td>Al-Wesabi, Choudhury, and Won (2018)</td>
<td>GNB, DT, LR, KNN and SVM</td>
<td>DT classifier developed with lesser features outperformed the other classifiers with an accuracy of 97%.</td>
<td>Model overfitting. Hold out method for splitting small sample dataset.</td>
</tr>
</tbody>
</table>
## Related work

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Techniques</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosgrave (2018)</td>
<td>SVM, random forest and GBM</td>
<td>GBM with fewer features and balanced class outperformed the other classifiers with a sensitivity of 77.8%.</td>
<td>Model performance can be improve.</td>
</tr>
<tr>
<td>Punjani and Atkotiya, 2019</td>
<td>DT</td>
<td>Developed an ML model that can predict whether a patient has cervical cancer or not.</td>
<td>There was no technique to handle class imbalance.</td>
</tr>
<tr>
<td>Ghoneim, Muhammad, and Hossain (2020)</td>
<td>Neural network</td>
<td>The neural network based model achieved an 91.2% accuracy</td>
<td>Neural network proves to be time consuming and requires huge datasets.</td>
</tr>
</tbody>
</table>
Methodology

1. Data Description and Visualization
2. Data Pre-processing
3. Classification
4. Model Evaluation
Data Description

- 858 samples
- 32 risk factors
- 803 Non-Cancerous
- 55 Cancerous
- **Biopsy**

*Source: Hospital 'Universitario de Caracas' in Caracas, the dataset is published on UCI machine learning repository in the year 2017.*
Methodology cont…

• **Data Preprocessing**
  • This involves missing values treatment using standard measurements like the mean for numerical attributes and mode categorical attributes.
  • Several patients decided not to answer some questions due to personal reasons. As a result, 13% of total questions were missed.
  • There are two features with 92% of missing values which are STDs: Time since first diagnosis and STDs: Time since last diagnosis, so they have to be omitted due to the high amount of incomplete values.
  • Rows with 20 or more missing values were also omitted.
Methodology cont…

- **Decision Tree (DT)**
- An intuitive model that makes decision based on a sequence of branching test.
- It can be describe as series of **yes or no** questions asked about feature values leading to predicted classes (for classification).
- Decision tree is easy to understand and interpret since it can be visualized.
- Decision tree can handle both numerical and categorical data.
Methodology cont…

- **Method of Data Splitting**
- The dataset was divided into training and testing data respectively with 10-Fold Cross Validation.
- Kfold is usually applied to datasets with limited observations.
- To avoid building a Decision tree model that can lead to overfit.
Methodology cont...

- Apply RFE
- How the RFE technique works
- **Step 1:** Train the model using all features
- **Step 2:** Determine model’s accuracy
- **Step 3:** Determine feature’s importance to the model for each feature
- **Step 4:** for each *subset size* $S_i$, $i = 1 \ldots N$ do
  - **Step 4.1:** Keep the $S_i$ most important features
  - **Step 4.2:** Train the model using $S_i$ features
  - **Step 4.3:** Determine model’s accuracy
- **Step 5:** end for
- **Step 6:** Calculate the accuracy profile over the $S_i$
- **Step 7:** Determine the appropriate number of features
- **Step 8:** Use the model corresponding to the optimal $S_i$
Methodology cont...

- Relevant features selected by RFE

<table>
<thead>
<tr>
<th>S/N</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
</tr>
<tr>
<td>2</td>
<td>Dx:</td>
</tr>
<tr>
<td>3</td>
<td>Dx: CIN</td>
</tr>
<tr>
<td>4</td>
<td>Dx: Cancer</td>
</tr>
<tr>
<td>5</td>
<td>STDs: Number of diagnosis</td>
</tr>
<tr>
<td>6</td>
<td>STDs: HPV</td>
</tr>
<tr>
<td>7</td>
<td>STDs: Hepatitis B</td>
</tr>
<tr>
<td>8</td>
<td>STDs: HIV</td>
</tr>
<tr>
<td>9</td>
<td>STDs: AIDS</td>
</tr>
<tr>
<td>10</td>
<td>STDs: molluscum contagiosum</td>
</tr>
<tr>
<td>11</td>
<td>Hormonal Contraceptives (Years)</td>
</tr>
<tr>
<td>12</td>
<td>Number of Sexual Partners</td>
</tr>
<tr>
<td>13</td>
<td>First Sexual Intercourse</td>
</tr>
<tr>
<td>14</td>
<td>STDs</td>
</tr>
<tr>
<td>15</td>
<td>IUD (Years)</td>
</tr>
<tr>
<td>16</td>
<td>IUD</td>
</tr>
<tr>
<td>17</td>
<td>Hormonal Contraceptives</td>
</tr>
<tr>
<td>18</td>
<td>Smokes (pack/year)</td>
</tr>
<tr>
<td>19</td>
<td>Smokes</td>
</tr>
<tr>
<td>20</td>
<td>Number of pregnancies</td>
</tr>
</tbody>
</table>
Methodology cont…

**Apply LASSO Regularisation**

- LASSO is a linear model that is simple and useful to use.
- The rationale behind LASSO technique is to fit the LASSO regression model on our dataset and consider only those features with coefficient greater than 0.
- One major advantage of LASSO is that it performs automatic feature selection.
- LASSO transforms each and every coefficient by constant component $\lambda$.
- If there is high correlation in the group of features, LASSO chooses only one among them and shrinks the others to zero. It minimizes the variability of the estimates by shrinking some of the coefficients exactly to zero.
Methodology cont...

- Relevant features selected by LASSO

<table>
<thead>
<tr>
<th>Features</th>
<th>LASSO Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-2.912</td>
</tr>
<tr>
<td>Number of sexual partners</td>
<td>-4.133</td>
</tr>
<tr>
<td>First sexual intercourse (age)</td>
<td>-6.216</td>
</tr>
<tr>
<td>Smokes (years)</td>
<td>-2.697</td>
</tr>
<tr>
<td>Smokes (packs/year)</td>
<td>1.858</td>
</tr>
<tr>
<td>Hormonal contraceptives (years)</td>
<td>5.706</td>
</tr>
<tr>
<td>STDs</td>
<td>1.802</td>
</tr>
<tr>
<td>STDs: Syphilis</td>
<td>-8.344</td>
</tr>
<tr>
<td>Dx: CIN</td>
<td>2.473</td>
</tr>
<tr>
<td>Dx:</td>
<td>2.875</td>
</tr>
</tbody>
</table>
Evaluation Metrics

- Some notations are used to present the results of our findings. These notations are defined as follows:

1. Accuracy: It is the number of correct predictions by the model out of the total number of observations as shown in equation (1).

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)
\]

2. Sensitivity: Is the ability of the classifier to correctly identify the positive class (cancerous). It is defined mathematically in equation (2).

\[
Sensitivity = \frac{TP}{TP + FP} \quad (2)
\]

3. Specificity: Is the percentage of times the classifier predicted a Negative class out of all the times the class was is Negative (non-cancerous). Equation (3) give the definition of specificity

\[
Specificity = \frac{TN}{FP + TN} \quad (3)
\]
Some notations are used to present the results of our findings. These notations are defined as follows:

4. Precision: Precision measures the ratio of people with cervical cancer and are correctly predicted by the model. It is defined mathematically in equation (4).

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

(4)

5. ROC Curve: The ROC is a technique used for visualizing the CCPM’s performance. The area under the ROC curve, abbreviated as AUC is used to evaluate the CCPM’s performance. The larger the AUC, the better the performance of the CCPM.
# Results

<table>
<thead>
<tr>
<th>Evaluation Metrics</th>
<th>DT Performance (All features)</th>
<th>DT + RFE Performance (20 features)</th>
<th>DT + LASSO (10 features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.9529</td>
<td>0.9765</td>
<td>0.9647</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.8571</td>
<td>0.8571</td>
<td>0.71429</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9615</td>
<td>0.9872</td>
<td>0.9872</td>
</tr>
<tr>
<td>Precision</td>
<td>0.6666</td>
<td>0.8571</td>
<td>0.8333</td>
</tr>
<tr>
<td>AUC</td>
<td>0.7446</td>
<td>0.9154</td>
<td>0.9435</td>
</tr>
</tbody>
</table>
Results Cont...

ROC curve for classifier without feature selection
Results Cont...

ROC curve for DT with RFE

ROC curve for DT with LASSO
Research Architecture
Contribution to research

- The contribution of this research are as follows:
- Applied RFE to select most relevant features for the classification of cervical cancer.
- Applied LASSO regularization method to select most relevant features in the prediction of cervical cancer.
The present study was designed to develop a model for the prediction of cervical cancer named a novel machine learning model for the classification of cervical cancer using Decision Tree with a Recursive Feature Elimination.

The model can be used as a decision support tool for the early detection of cervical cancer in patients.

The model will enable clinical practitioners to isolate and manage patients who show early sign of cervical cancer.
Future Work

- The issue of predicting cervical at an early stage can be usefully explored in further research.
- Further research regarding the role of other feature selection methods would be worthwhile.
- It is recommended that further research be undertaken with a very large volume of the dataset so that in-depth analysis and understanding can be performed and a better predictive model can be developed for the same problem.
- Future work is intended to be carried out for the long term management of cervical cancer in clinical and personalized medical management by decision support system.
References

References cont…

THANK YOU FOR YOUR ATTENTION