Project Report

Cervical Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques

Machine Learning (CS5439)

By:
1806064 – Anant Malhotra

Submitted to:

Dr. Akshay Deepak

Department of Computer Science and Engineering,

NIT Patna

REFERENCES

Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques

SHERIF F. ABDOH , MOHAMED ABO RIZKA, AND FAHIMA A. MAGHRABY
Department of Computer Science, Arab Academy for Science, Technology and Maritime Transport, Cairo 1029, Egypt
Corresponding author: Sherif F. Abdoh (sherif.fayz@outlook.com)

IEEE ACCESS (ISSN / eISSN : 2169-3536)
Received August 29, 2018, accepted September 26, 2018, date of publication October 5, 2018, date of current version October 31, 2018.
Digital Object Identifier 10.1109/ACCESS.2018.2874063

INTRODUCTION

Cervical cancer is a serious worldwide health problem. The percentage of cervical cancer cases in developing countries is 80%. The United States estimate 13.240 new cervical cancer cases in 2018 and about 4.170 estimated death which means that the death ratio is nearly 31.5%. This type of cancer affects the female reproductive system by attacking women's cervix area. In most cases, it grows without any symptoms at its early stages. The symptoms appear at its late stages which make it hard to be cured and the disease may spread to other organs of the body. That's why its diagnosis at its early stages is very important to improve its cure and survival ratios.

In this project Synthetic Minority Oversampling Technique (SMOTE) algorithm is used to balance the dataset classes by increasing the number of the minority class based on k-nearest neighbours to nearly equal classes. In addition, Recursive Feature Elimination (RFE) and Principle Component Analysis (PCA) are used as feature reduction techniques to reduce the processing time and to neglect unimportant features from being used in the classification. Then RF classification technique is used to classify the cases into cervical cancer and non-cervical ones. Finally, the model performance is measured before and after SMOTE then compared with other related work results.

CERVICAL CANCER DATASET

- The used dataset was published on the repository of **University of California at Irvine (UCI)** collected at Hospital Universitario de Caracas in Caracas, Venezuela.
- The dataset comprises historical medical records, habits and demographic information for **858** cases with **32** features for each case.
- The features - Age, STDs: Number of diagnosis, Dx:Cancer, Dx:CIN, Dx:HPV, Dx are of Integer type; rest all are of String type.
- 4 target variables: Hinselmann, Schiller, Cytology and Biopsy.
- From given table it is found that there are a lot of missing values in features 27 and 28, so feature 27 and 28 are removed.

Number	Features	Entries	Missing data
1	Age	858	0
2	Number of sexual partners	832	26
3	First sexual intercourse	851	7
4	Num of pregnancies	802	56
5	Smokes	845	13
6	Smokes (years)	845	13
7	Smokes (packs/year)	845	13
8	Hormonal Contraceptives	750	108
9	Hormonal Contraceptives (years)	750	108
10	IUD	741	117
11	IUD (years)	741	117
12	STDs	753	105
13	STDs (number)	753	105
14	STDs:condylomatosis	753	105
15	STDs:cervical condylomatosis	753	105
16	STDs:vaginal condylomatosis	753	105
17	STDs:vulvo-perineal condylomatosis	753	105
18	STDs:syphilis	753	105
19	STDs:pelvic inflammatory disease	753	105
20	STDs:genital herpes	753	105
21	STDs:molluscum contagiosum	753	105
22	STDs:AIDS	753	105
23	STDs:HIV	753	105
24	STDs:Hepatitis B	753	105
25	STDs:HPV	753	105
26	STDs: Number of diagnosis	858	0
27	STDs: Time since first diagnosis	71	787
28	STDs: Time since last diagnosis	71	787
29	Dx:Cancer	858	0
30	Dx:CIN	858	0
31	Dx:HPV	858	0
32	Dx	858	0

PRE-PROCESSING AND SMOTE OVERSAMPLING

- The mean equation is used to handle the missing values.

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Dataset is unbalanced as the number of negative class records is larger than the positive class. So, SMOTE oversampling technique is used to solve the unbalanced problems.
- SMOTE is used to synthetically increase the minority class based on k-nearest neighbours.
- SMOTE technique uses the following equation to synthetically increase the minority class.

$$x_{syn} = x_i + (x_{knn} - x_i) \times t$$

- The given table shows that before SMOTE the dataset was imbalanced and after the implementation of SMOTE algorithm it ends up with almost balancing the dataset.

Number of records before and after SMOTE.

Examination	Before SMOTE		After SMOTE	
	Patient	*Non-patient*	*Patient*	*Non-patient*
Hinselmann	35	823	805	823
Schiller	74	784	740	784
Citology	44	814	792	814
Biopsy	55	803	770	803

SAMPLE ROW BEFORE & AFTER PREPROCESSING AND SCALING

Age	18
Number of sexual partners	4.0
First sexual intercourse	15.0
Num of pregnancies	1.0
Smokes	0.0
Smokes (years)	0.0
Smokes (packs/year)	0.0
Hormonal Contraceptives	0.0
Hormonal Contraceptives (years)	0.0
IUD	0.0
IUD (years)	0.0
STDs	0.0
STDs (number)	0.0
STDs:condylomatosis	0.0
STDs:cervical condylomatosis	0.0
STDs:vaginal condylomatosis	0.0
STDs:vulvo-perineal condylomatosis	0.0
STDs:syphilis	0.0
STDs:pelvic inflammatory disease	0.0
STDs:genital herpes	0.0
STDs:molluscum contagiosum	0.0
STDs:AIDS	0.0
STDs:HIV	0.0
STDs:Hepatitis B	0.0
STDs:HPV	0.0
STDs: Number of diagnosis	0
STDs: Time since first diagnosis	7
STDs: Time since last diagnosis	7
Dx:Cancer	0
Dx:CIN	0
Dx:HPV	0
Dx	0

Age	-1.038563
Number of sexual partners	0.897061
First sexual intercourse	-0.715096
Num of pregnancies	-0.912086
Smokes	-0.415910
Smokes (years)	-0.300756
Smokes (packs/year)	-0.205194
Hormonal Contraceptives	-1.430242
Hormonal Contraceptives (years)	-0.641569
IUD	-0.382174
IUD (years)	-0.285284
STDs	-0.365452
STDs (number)	-0.335707
STDs:condylomatosis	-0.265919
STDs:cervical condylomatosis	0.000000
STDs:vaginal condylomatosis	-0.078007
STDs:vulvo-perineal condylomatosis	-0.262695
STDs:syphilis	-0.167047
STDs:pelvic inflammatory disease	-0.038926
STDs:genital herpes	-0.038926
STDs:molluscum contagiosum	-0.038926
STDs:AIDS	0.000000
STDs:HIV	-0.167047
STDs:Hepatitis B	-0.038926
STDs:HPV	-0.055086
STDs: Number of diagnosis	-0.289093
Dx:Cancer	-0.146385
Dx:CIN	-0.102960
Dx:HPV	-0.146385
Dx	-0.169630

FEATURE SELECTION

The model uses two feature selection (dimensionality reduction) methods which are Principle Component Analysis (PCA) and Recursive Feature Elimination (RFE).

PCA

The main idea of PCA is to map the n-dimension feature space into k-dimension which is also known as principle component where k < n.

The covariance matrix is computed whereby the result is used to calculate the eigenvectors and eigen values. The eigen vector with the highest eigen value is chosen as the principle component of the dataset as it exhibits the most significant relationship between the data set attributes.

$$A = \mathrm{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj})$$

Eigenvalue

$$A\mathbf{x} = \lambda\mathbf{x}$$

Eigenvector

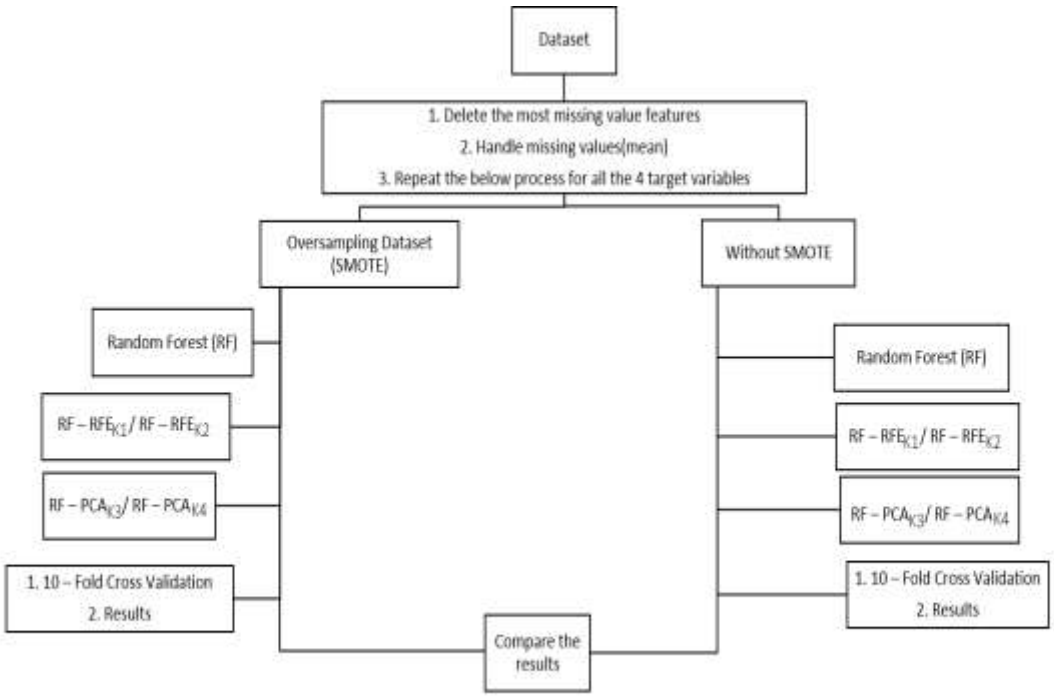
RFE

RFE is basically a backward selection of the predictors. This technique begins by building a model on the entire set of predictors and computing an importance score for each predictor. The least important predictor(s) are then removed, the model is re-built, and importance scores are computed again.

Input:
Data set T
Set of p original features F = {f1,f2,..,f p}
Output:
Subset of features
Code:
Final ranking R
Repeat for i in {1: p − 1}
Rank set F using random forest
f* ← last ranked feature in F
R(p −i +1) ← f
* F ← F − f*

PROPOSED MODEL

- The main idea of our model is to diagnose cervical cancer using random forest with SMOTE and two feature reduction techniques.
- The variables K1, K2, K3 and K4 represent the number of selected features. (Their values are taken from the paper)



RANDOM FOREST CLASSIFIER

- Random Forest (RF) is a **supervised** classification technique that works on the principle of using group of weak learners to form a strong learner.
- The construction of RF can be described in the following steps.

1. Generates N number of bootstrap samples from the dataset.
2. Each node takes a random sample of attributes of size m where m < M. (M refers to the total number of attributes).

3. Constructs a split using the m attributes selected in Step 2 and calculates the k node using the best split point. (‘‘k’’ refer to next node).
4. Repeats splitting the tree until only 1 leaf node is reached and the tree is completed.
5. The algorithm is trained on each bootstrapped separately.
6. Uses the trees classification voting to collect the prediction data from the (n) trained trees.
7. Uses the highest voted features to build the final RF model.

PERFORMANCE COMPARISON (PROPOSED vs IMPLEMENTED)

Hinselmann (Proposed Model)										
	RF	RF-RFE		RF-PCA		SMOTE-RF	SMOTE-RF-RFE		SMOTE-RF-PCA	
#Features	30	5	15	5	11	30	5	15	5	11
Accuracy	95.92	95.33	95.80	96.03	95.92	97.60	95.14	95.88	97.42	97.48
Hinselmann (Implemented Model)										
Accuracy	95.68	95.45	95.68	95.57	95.45	97.39	96.41	97.45	89.97	94.10

Schiller (Proposed Model)										
	RF	RF-RFE		RF-PCA		SMOTE-RF	SMOTE-RF-RFE		SMOTE-RF-PCA	
#Features	30	7	18	6	12	30	7	18	6	12
Accuracy	91.49	90.79	91.25	90.56	90.91	95.01	91.73	92.91	94.49	94.88
Schiller (Implemented Model)										
Accuracy	90.55	90.79	90.32	90.09	90.20	94.83	94.32	94.58	89.41	92.92

Citology (Proposed Model)										
	RF	RF-RFE		RF-PCA		SMOTE-RF	SMOTE-RF-RFE		SMOTE-RF-PCA	
#Features	30	8	15	8	11	30	8	15	8	11
Accuracy	94.52	93.47	94.17	94.52	94.63	96.94	92.52	95.89	96.39	96.89
Citology (Implemented Model)										
Accuracy	94.40	94.87	94.52	94.75	94.40	96.00	94.96	95.94	92.25	92.93

Biopsy (Proposed Model)										
	RF	RF-RFE		RF-PCA		SMOTE-RF	SMOTE-RF-RFE		SMOTE-RF-PCA	
#Features	30	6	18	8	11	30	6	18	8	11
Accuracy	93.70	93.12	93.24	93.24	93.24	96.06	95.23	95.87	95.55	95.74
Biopsy (Implemented Model)										
Accuracy	93.70	93.24	93.35	93.47	93.12	97.01	95.58	96.32	91.90	93.34