Different Classification Techniques on Cervical Cancer Data

Sangita Mitra, Rejwana Haque, Afia Anjum
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
Email:{sangita, rejwana1, afia}@ualberta.ca

***Abstract*—This Cervical cancer is one of the most common gynecological cancer around the world. In our approach, we look into finding the best screening strategy for a cervical cancer patient by applying three different machine learning algorithms namely logistic regression, support vector machine and neural network. These classification techniques work on thirty-two different features from eight hundred and fifty-eight samples which initiate the risk factor of cervical cancer. Two significant impediments of this data are missing value and imbalance data. Therefore, in our experiment, we use mutual imputation to handle missing value as well as over-sampling and Synthetic Minority Oversampling Technique (SMOTE) approach to balance the data. Our experiment shows that neural network is the best classifier with 85.35% accuracy on this dataset.**

Keywords—Cervical Cancer, Neural Network, Support Vector Machine, Logistic Regression

I. INTRODUCTION

Every year, more than half a million women are diagnosed with cervical cancer and over three lakh women died suffering from this cancer worldwide [1]. Cervical cancer has become a major health challenge around the world as it is the fourth most frequent cancer in women with an estimated 570,000 new cases in 2018 representing 6.6% of all female cancers. The rate of cervical cancer is more prevalent among low income and middle-income countries where mortality is 18 times higher than that in developed countries [2]. Approximately 90% of deaths from cervical cancer occurred in low and middle-income countries. In 2012, 85% of cervical cancers occur in less-developed regions with the diagnosis of 530000 cervical cancer cases approximately [3]. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide.

that The high mortality rate from cervical cancer globally could be reduced through a comprehensive approach that includes prevention, early diagnosis, effective screening and treatment program. In fact, the World Health Organization (WHO), called for action towards achieving the global elimination of cervical cancer through improving HPV vaccination, cervical screening and pre-cancer and cancer treatment, particularly in low and middle-income countries [4]. For this purpose, cervical cancer dataset can be an effective way of predicting risk factors of cervical cancer and deciding best screening strategy.

We are addressing the research question: What model and hyper-parameters are best to decide whether biopsy test should use or not for an individual cervical cancer patient? We use logistic regression, support vector machine and neural network in this purpose.

II. BACKGROUND AND RELATED WORK

Cervical cancer is a type of cancer which happens due to abnormal growth of cell in the cervix part of human body. It is possible to recover from cervical cancer if it can be diagnosed early. For this purpose, [5] developed a novel bio-marker from specific proteins, enzymes and metabolites. Pap smear test was used as an effective screening method to determine the different stages of cervical cancer. Both geometric and texture features of Pap smear images have been used for cervical cancer detection and for classification three different kernel SVM have implemented in [6].

Cervical cancer data has been studied from different researchers in last few years. The cervical cancer dataset [7] has been published in 2017 including 858 samples and 32 features as well as four targets. Because of several screening and diagnosis procedures in cervical cancer, it leads to a complex ecosystem [8] is the first work on this dataset which proposes a regularization-based transfer learning strategy to transfer the contribution type of each feature on linear models to show its impact on sharping accuracy. Three support vector machine-based approaches have been used on this dataset to diagnosis cervical cancer where SVM Principal Component Analysis achieved 94% accuracy [9]. confirm that you have the correct template for your paper size.

III. DATASET DESCRIPTION

Cervical cancer dataset was collected at ‘Hospital Universitario de Caracas’ in Caracas, Venezuela. The multivariate dataset comprises demographic information, habits like smoking, and historic medical records of 858 patients. The attributes and their types are given in Table:1. The target values are four different test

TABLE I. ATTRIBUTES AND THEIR TYPES

Attribute	Type	Attribute	Type	Attribute	Type
Age	Integer	STDs	Bool	STDs:HIV	Bool
Number of sexual partners	Integer	STDs (number)	Integer	STDs:Hepatitis B	Bool
First sexual intercourse (age)	Integer	STDs:condylomatosis	Bool	STDs:HPV	Bool
Number of pregnancies	Integer	STDs:cervical condylomatosis	Bool	STDs: Number of diagnosis	Integer
Smokes	Bool	STDs:vaginal condylomatosis	Bool	STDs: Time since first diagnosis	Integer
Smokes (years)	Bool	STDs:vulvo-perineal condylomatosis	Bool	STDs: Time since last diagnosis	Integer
Smokes (packs/year)	Bool	STDs:syphilis	Bool	Dx:Cancer	Bool
Hormonal Contraceptives	Bool	STDs:pelvic inflammatory disease	Bool	Dx:CIN	Bool
Hormonal Contraceptives (years)	Integer	STDs:genital herpes	Bool	Dx:HPV	Bool
IUD	Bool	STDs:molluscum contagiosum	Bool	Dx	Bool
IUD (years)	Integer	STDs:AIDS	Bool		

results used to find the risk factor of cervical cancer: Hinselmann, Schiller, Cytology and Biopsy. This study is going to focus on the Biopsy target for comparing the performance of the algorithms in terms of percentage accuracy.

All 32 features are not equally important to calculate the risk factors of cervical cancer as [10] shows that out of 32 different features, six (age, first sexual intercourse, number of pregnancies,smokes, hormonal contraceptives and STDs: genital herpes) are the main predictive features that play a vital role in detecting the risk factor for cervical cancer. This research comes to this conclusion by showing that they have an accuracy of 97.5% using the 6 main features and utilizing Decision Tree Classifier. Some dataset related terms are: UID (Intrauterine Contraceptive Device), HPV(Human Papilloma Virus), HIV(Human Immunodeficiency Virus), CIN(Cervical Intraepithelial Neoplasia), STD(Sexually Transmitted Diseases).

IV. METHODOLOGY

A. Algorithms and parameters

We are implementing three different algorithms (Logistic Regression, Support Vector Machine and Neural Network) with different parameters in order to identify the risk of cervical cancer of an individual based on his/her biopsy result. For implementation we use scikit-learn tools in python.

1) Logistic Regression: Logistic regression technique is a widely accepted technique for statistical analysis and modelling where the probability of the outcome is dependent on the series of predictor variables [11]. There are different types of Logistic Regression such as the binary logistic regression, multinomial logistic regression, ordinal logistic regression, etc. Since in this dataset, our target (risk of cervical cancer based on biopsy result) is categorical, we will be using Binary Logistic Regression, which is based on Bernoulli distribution. The input data pass through a prediction function and returns a probability score between 0 and 1. We use sigmoid and tanh transfer function as in machine learning, these functions use widely to map predictions to probabilities. In our tests, the threshold value varies from 0.5 to 0.9 by skipping 0.1. Stochastic Gradient Descent (SGD) epoch will be used as it performs parameter update for each observation.

2) Support Vector Machine: SVM (Support Vector Machine) maps data into a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. That is, it can solve linear and non-linear problems and work well for many practical problems. It finds a separating line between different categories and then the data are transformed in such a way that the separator could be drawn as a hyper-plane. These Hyper-plane are decision boundaries that help to classify the data points.

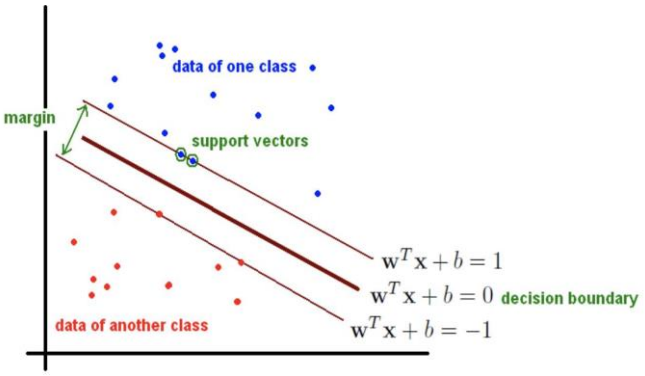

Fig. 1. Support vector machine.

Data points falling on either side of the hyper-plane can be attributed to different classes. Also, the dimension of the hyper-plane depends upon the number of features. We could use different kernel functions for this transformation such as Linear, Polynomial, Radial Basis Function and Sigmoid. Previous works on this dataset justified that SVM is an effective algorithm for classifying the target value.

3) Neural Network: A neural network is a set of layers of highly interconnected processing elements that make a series of transformations on the data to generate its own understanding of features. It is a universal function approximator and it can find representations for the input in different dimensioned space. A neural network has 3 types of layers: input layer, hidden layer and output layer. There is an activation function for transferring information from one layer to another layer. After calculation the loss we will update the weights of every layer backwards to minimize the difference, get it as close to zero as possible. In this way, the weight at each layer will be updated iteratively by neural network. In our tests, we use a differing number of hidden layers from 1 to 4, varying the number of nodes between 8 to 64 in each layer. We use three different activation function namely sigmoid, ReLU and tanh.

B. Experiments

The experiment will be performed in the following three steps.

a) Data pre-processing: One of the constraints of this dataset is that several patients decided not to answer some of the questions because of privacy concerns, that's why this

data suffers from missing values. This dataset includes 3,622 missing values out of 27,456 observations, which forms 13.2% of the data [12]. But for predicting cervical cancer risk factor, it is very important to get that personal information. The missing values in a dataset may generate bias, affecting the quality of the supervised learning process or the performance of classification algorithms. That's why it is important to handle with missing value in this study.

For this scope of the study, we will be considering the technique of multiple imputation to deal with missing values [13]. This technique takes advantage of the correlation between responses. As multiple imputation produces appropriate results even in the presence of small sample size or a high number of missing data, it is a reasonable choice to use multiple imputation for handling missing values in our experimental setting [14].

This dataset also suffers from imbalanced data that means the classification categories are not approximately equally represented in the dataset results into biased data [15]. To address the biased data, there are several methods available like SMOTE, under-sampling and over-sampling method. As our dataset is small, oversampling will be helpful to balance that dataset and we use the SMOTE [16] method as it works significantly good for multi-class classification. The SMOTE algorithm carries out an oversampling approach to rebalance the original training set. Instead of applying a simple replication of the minority class instances, the key idea of SMOTE is to introduce synthetic examples. This new data is created by interpolation between several minority class instances that are within a defined neighborhood.

b) Learning models and selecting the best hyper-parameters: After getting the pre processed data, the system learns three different models and by using 10-fold cross validation and then compute the average error of each of the hyper-parameters. In this way, we get the best hyper-parameter for each algorithm.

c) K-fold cross-validation for measuring accuracy: Using the hyper-parameter selected by K fold cross-validation, we further run our algorithms multiple times using separate train-test dataset. As there is only one dataset on cervical cancer, we randomly sub-sample the dataset with the splitting factor of 0.8 over multiple runs and report the error. For each run, we train on 7 split and test on last split. In this phase, we apply the k-fold cross-validation 3 times that gives accuracy for individual algorithm.

V. Analysis

Confusion matric is a simple and effective way for classification problem to analyze the performance of a model. To compare our algorithms, we calculate the following confusion matrix in Table II: accuracy, recall, precision, and F1 score. The accuracy of the model is basically the total number of correct predictions divided by total number of predictions. The precision of a class define how credible is the result when the model answer that a point belongs to that class. The recall of a class expresses how well the model is able to detect that class. The F1 score of a class is given by the harmonic mean of precision and recall (2×precision×recall / (precision + recall)), it combines precision and recall of a class in one

$$\text{Accuracy} = TP+TN/TP+FP+FN+TN \quad (1)$$

$$\text{Precision} = TP/TP+FP \quad (2)$$

$$\text{Recall}=TP/TP+FN \quad (3)$$

metric. ROC curve is a another performance measurement for classification problem at various thresholds settings. Based on the ROC curve, we can build another metric, easier to use, to evaluate the model: the Area Under the ROC curve(AUROC). AUROC acts a little bit as a scalar value that summarizes the entire ROC curve. From our experiment, we get 95% accuracy for three algorithms as our dataset are suffering from imbalanced data. After using the SMOTE approach, this problem has solved and we get 69.84%, 72.16% and accuracy for three individual algorithm.We use Binomial test on these measures, to test our hypothesis of

TABLE II. STATISTICAL SCORES

Algorithms	Accuracy(%)	Recall(%)	F1
Logistic Regression	95.128	1.002	0
SVM	95.867	1.112	0
NN	95.043	1.011	0

TABLE III. Classifier Performance Using SMOTE Method.

Algorithms	Accuracy	Recall	precision	F1
Logistic Regression	73.841	71.184	75.445	70.246
SVM	75.161	87.905	67.326	74.891
NN	85.345	89.231	82.197	84.218

which is the best algorithm having lower performance error

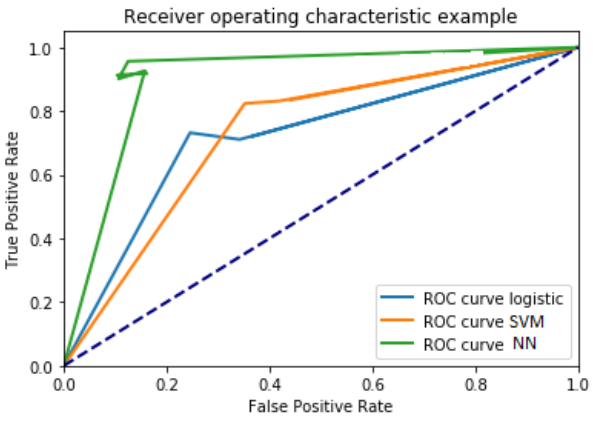

Fig.2. ROC curve.

(from step3) is awarded a win for that specific run using equation (4). For each of the following combinations of the algorithms we will perform Binomial test (with significance level of 0.05):

1) SVM vs. Logistic Regression
2) SVM vs. Neural Network

In the above three settings, the null hypothesis (Ho) will be that the quality of both algorithm are same, while the alternative hypothesis (Ha) will be that, one of the algorithms performs

$$P = \sum_{i=k}^{m} \binom{m}{i} p^i (1-p)^{m-i} \quad (4)$$

better than the other. We have used SVM as baseline in this significance test. Finally, we will select the winner of the three algorithms for our problem.

VI. Discussion and Conclusion

This paper shows a comparative analysis of three different machine learning classifiers on cervical cancer dataset and find the best model for selecting effective screening strategy for diagnosis of cervical cancer. As the dataset suffers from imbalance data and missing value, that's why, SMOTE approach for over-sampling and mutual imputation have been used to solve this problem. Eventually, neural network is the winner algorithm. Our research has a few limitations, limiting its generalizability. In our approach, we use only Biopsy test as target value although all four tests are important to diagnosis the cervical cancer more accurately. In future work, we will do multi-class classification targeting four screening test for cervical cancer.

References

[1] Cohen, P. A., Jhingran, A., Oaknin, A., & Denny, L. (2019). "Cervical cancer". The Lancet, 393(10167), 169-182. DOI: 10.1016/S0140-6736(18)32470-X.

[2] WHO. Cervical cancer. World Health Organization https://www.who.int/cancer/cervical-cancer.

[3] Simms, K. T., Steinberg, J., Caruana, M., Smith, M. A., Lew, J. B., Soerjomataram, I., & Canfell, K. (2019). "Impact of scaled up human papillomavirus vaccination and cervical screening and the potential for global elimination of cervical cancer in 181 countries, 2020–99: a modelling study." The Lancet Oncology, 20(3), 394-407. DOI: 10.1016/S1470-2045(18)30836-2.

[4] Canfell, K. (2019). "Towards the global elimination of cervical cancer". Papillomavirus Research, 100170. DOI: 10.1016/j.pvr.2019.100170.

[5] Dasari, S., Wudayagiri, R., & Valluru, L. (2015). "Cervical cancer: Biomarkers for diagnosis and treatment". Clinica chimica acta, 445, 7-11. DOI:10.1016/j.cca.2015.03.005.

[6] Kashyap, D., Somani, A., Shekhar, J., Bhan, A., Dutta, M. K., Burget, R., & Riha, K. (2016, June). "Cervical cancer detection and classification using Independent Level sets and multi SVMs". In 2016 39th international conference on telecommunications and signal processing (TSP) (pp. 523-528). IEEE. DOI: 10.1109/TSP.2016.7760935.

[7] https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. Cervical cancer dataset.

[8] Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017, June). "Transfer learning with partial observability applied to cervical cancer screening". In Iberian conference on pattern recognition and image analysis (pp. 243-250). Springer, Cham. DOI:10.1007/978-3-319-58838-4_27.

[9] Wu, W., & Zhou, H. (2017). "Data-driven diagnosis of cervical cancer with support vector machine-based approaches". IEEE Access, 5, 25189-25195.DOI: 10.1109/ACCESS.2017.2763984.

[10] Al-Wesabi, Y. M. S., Choudhury, A., & Won, D. (2018, December). "Classification of cervical cancer dataset". Inn Avishek Choudhury, Wesabi, Classification of Cervical Cancer Dataset, Proceedings of the 2018 IISE Annual Conference, Orlando (pp. 1456-1461).

[11] Singh, S., Panday, S., Panday, M., & Rautaray, S. S. (2019). "Logistic Regression for the Diagnosis of Cervical Cancer". In Data, Engineering and Applications (pp. 109-117). Springer, Singapore. DOI: 10.1007/978-981-13-6347-4_10.

[12] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). "A gentle introduction to imputation of missing values". Journal of clinical epidemiology", 59(10), 1087-1091. DOI: 10.1016/j.jclinepi.2006.01.014.

[13] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). "Missing data imputation using statistical and machine learning methods in a real breast cancer problem". Artificial intelligence in medicine, 50(2), 105-115. DOI: 10.1016/j.artmed.2010.05.002.

[14] Chawla, N. V. (2009). "Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook" (pp. 875-886). Springer, Boston, MA.

[15] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary". Journal of artificial intelligence research, 61, 863-905. DOI: 10.1613/jair.1.11192.