International Journal of Computer Applications (0975 - 8887)
Volume 182 - No.46, March 2019

Analyzing Cervical Cancer by using an Ensemble Learning Approach based on Meta Classifier

Lipika Barua
Department of CSE
Gono Bishwabidyalay
Savar, Dhaka

Md. Sharif Ahamed
Department of CSE
Gono Bishwabidyalay
Savar, Dhaka

Tania Akter
Department of CSE
Jahangirnagar University
Savar, Dhaka

ABSTRACT

Now a days, cervical cancer is treated as one of the main causes of death of women due to cancer in worldwide. In this study, we collected 741 instances of cervical cancer data, preprocessed data, explored high ranked significant features using different feature selection techniques such as Information Gain (Info. Gain), Gain Ratio, Gini Indexing and χ^2 and implemented different meta classifier techniques such as Dagging, Additive Regression, CVParameter Selection, MultiScheme, MultiSearch. After that we proposed a new ensemble learning method which is combined with different meta classifier algorithms. Then we found out different evaluation metrics such as Mean Absolute Logarithmic Error (MALE), Root Mean Square Logarithmic Error (RMSLE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) using each of meta classifier algorithms. Later we compared the result of error rate of an proposed algorithm with the error rate result of different meta classifier algorithms and it showed that the proposed algorithm performs as the best classifier to classify these cervical cancer data. This analysis will be helpful to evaluate a performance model for researcher.

Keywords

Cervical Cancer, Feature Selection, Feature Ranking, Classification, Ensemble learning method

1. INTRODUCTION

Cervical cancer is a cancer which arises from the cervix. It arises because of the abnormal growth of cells and then it extends over all other parts of the body. Today cervical cancer is considered as the second most common cause of cancer among women worldwide [12]. The burden of the disease is increasing day by day due to the ascending trend of transmissible diseases such as HIV and Human 67 Papilloma Virus (HPV). Several studies show that cervical cancer screening and intervention programs should target communities with lower socioeconomic status due to lower rates of screening and knowledge. According to WHO approximately 90% of deaths from cervical cancer occurred in low- and middle-income countries.Smoking is considered as one of the top most main causes for cervical cancer. Long-term use of oral contraceptive pills and also multiple pregnancies can also cause cervical cancer. In this present situation of world, overweight and obese women had an increased risk of cervical cancer, likely because of under-diagnoses of cervical pre-cancer [9]. In this paper, we have collected a data set of the factors that causes cervical cancer from UCI data repository. On the collected dataset, we have applied Correlation-based Feature Selection (CFS) methods with many subset classifiers. As well as we have evaluated high ranked significant features with applying different feature selection techniques such as Information Gain (Info. Gain), Gain Ratio, Gini Indexing and χ^2. Then we have calculated Mean Absolute Logarithmic Error (MALE), Root Mean Square Logarithmic Error (RMSLE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) using different meta classifier algorithms which gives the error rate.The classifier that shows the lowest error rate performs better as classifier. In this work, we have also proposed a new method based on ensemble learning approach. Then we compared its error rate for better performance with others meta classifier algorithms.

This paper is organized into several sections. Section II represents materials and methods where analysis of data and working procedure are described in brief. Section III represents experimental results of our work. Thus, various results on our findings are warranted in section IV. Finally, our experimental analysis are shortened with some limitations and finished this work with expecting some future plans.

2. RELATED WORK

In 2017, K. Fernandes, et al. [3] did a research to moderate the amount of labeled data from each modality/expert. They proposed a regularization-based transfer learning strategy that encourages source and target models to share the same coefficient signs. They instantiated the proposed framework to predict cross-modality individual risk and cross-expert subjective quality assessment of colonoscopy images for different modalities. In 2013, C.T. Tseng et al. [14] did a study where they applied machine learning techniques to produce objective to an inferential problem of recurrent cervical cancer. In this study, the machine learning approaches including support vector machine, C5.0 and extreme learning machine were considered to find important risk factors to predict the recurrence prognoses for cervical cancer. In 2010, X.Hu et al. [7] identified a microRNA based signature for the prediction of cervical cancer survival in their work. They developed a logistic regression model which was developed based on these two microRNAs and the prognostic value of the model was subsequently validated with independent cervical cancers. In 2006, K.Thanganel et al. [13] found how

the problem of cervical cancer diagnosis is approached by a data mining analyst with a background in machine learning. Here they made an attempt to identify patterns from the database of the cervical cancer patients using clustering. In 2004, S.Hu et al. [5] did a study where they identified the combined patterns of cervical cancer risk factors including demographic environmental and genetic factors using induction technique. They compared logistic regression and a decision tree algorithm and finally this study showed how the decision tree algorithm could be used in risk analysis and target segmentation for cervical cancer management. In 2004, J.T Horng et al. [6] in their study employed a Bayesian network and four decision tree algorithms and compared the performance of these learning algorithms. The result of this study showed the possibility of investigation that could identify combinations of genetic factors such as SNPs and microsatellites which influence the risk associated with common multifactorial diseases such as cervical cancer.
In 2004, Nathalie Reesink-Peters et al. [10] did a Current morphology-based cervical cancer screening which is associated with -significant false-positive and false-negative results. According to their study it is unknown whether a cervical scraping reflects the methylation status of the underlying epithelium and it is therefore unclear whether quantitative hyper methylation specific PCR (QMSP) on cervical scrapings could be used as a future screening method augmenting the current approach. This feasibility study showed that QMSP on cervical scrapings holds promise as a new diagnostic tool for cervical cancer. In 2003, R Sankaranarayanan, MD et al. [11] in their study they provide valuable information on the average, comparative test performances in detecting high-grade cervical cancer precursors and cancer. In 2001, Sue j. Goldie et al. [4] did a study to assess the cost-effectiveness of several cervical cancer screening strategies using population specific data. In their analysis they developed a comprehensive model capable of assessing alternative screening strategies for cervical cancer in developing countries and used the model together with country specific data to conduct a policy analysis comparing the clinical benefits and cost effectiveness of different cervical cancer screening strategies.

3. METHOD AND MATERIALS

We considered several steps to analyze cervical cancer data and find out significant ranked features using different feature selection methods and applied different meta classifier algorithms which are shown best performance to detect cancer which are given as follows:

3.1 Cervical Cancer Data

There were N=858 records which we collected from UCI machine learning data repository. The data set contains demographic information, habits and historic medical records of 858 patients. But there were some missing values among the records which were some of the questions that several patients did not answer because of privacy. After removing the missing values, we have selected 741 records for feature selection and classification approach. This cervical cancer dataset consists of 36 attributes where the outcome class is represented by binary value ”0” or ”1”.’ Here, ”1” value represents the positive cervical cancer and ”0” represents the negative cervical cancer.

Table 1. Feature Ranking.

	Info. Gain	**Gain Ratio**	**Gini**	χ^2
Schiller	0.059	0.127	0.015	94.221
Hinselmann	0.016	0.059	0.004	27.272
Dx:HPV	0.006	0.039	0.002	10.533
Dx:Cancer	0.005	0.035	0.001	9.592
Dx	0.003	0.025	0.001	7.338

3.2 Feature Selection and Ranking Approach

Feature selection is an important part in the field of data analysis. It help us to create an accurate predictive model by removing unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model. In this experiment, we have used many feature selection algorithms such as cfssubsetEvaluator with (AntSearch, CuckooSearch, BeeSearch, BatSearch, ElephantSearch, Evolutionary Search, FlowerSearch, GeneticSearch, FireflySearch, Greedy Stepwise, HarmonySearch, LinearForward Selection, Multiobjective Evolutionary Search, PSO Search, RhinocerosSearch, SubsetSize Forward Selection, WolfSearch). The original data records had 36 attributes with 4 class levels. After applying these different feature selection approaches, we have selected common 24 attributes for the different meta classification approach and selected ”cytology” as our class level.
We have also applied different feature selection approaches to evaluate and rank significant features of cervical cancer data. Different feature selection approach such as Information Gain (Info. Gain), Gain Ratio, Gini Indexing and χ^2 are represented in Table 1 . ”Schiller” feature shows as the high ranked feature and Hinselmann, Dx:HPV (Human Papilloma Virus), Dx:Cancer (person had previous cervical cancer diagnostic), Dx are also pointed high ranked features respectively according to Info. Gain, Gain Ratio, Gini Indexing and χ^2.

3.3 Classification Approach

Classication is the process to find function or model that explains and distinguishes concepts or classes whose label is unknown for the intention to predict the class of objects of using the model. After data preprocessing, we have used many meta classifier algorithms to analyze these datasets. But many of them were more complex and did not gave accurate outcomes. So finally we have selected five classification model (Dagging, Additive Regression, CVParameter Selection, MultiScheme, MultiSearch). Using these classifiers, we have found the MALE (Mean Absolute Logarithmic Error), RMSLE (Root Mean Squared Logarithmic Error), MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error). These values shows the different error rate result of the classifiers by which we do the performance analysis.Then we applied a proposed method and compared the performance with the previous results of the classifiers.

3.4 Proposed Algorithm

In this paper, we introduced a new proposed algorithm based on AttributeSelected meta classifier which combines with AdditiveRgression classifier to produce lower error rate and evaluate better performance of cervical cancer data and it supports for high-dimensional multi-class data using ensemble learning technique. Ensemble is an one kind of learning process which combined with different classification methods to generate a strong associ-

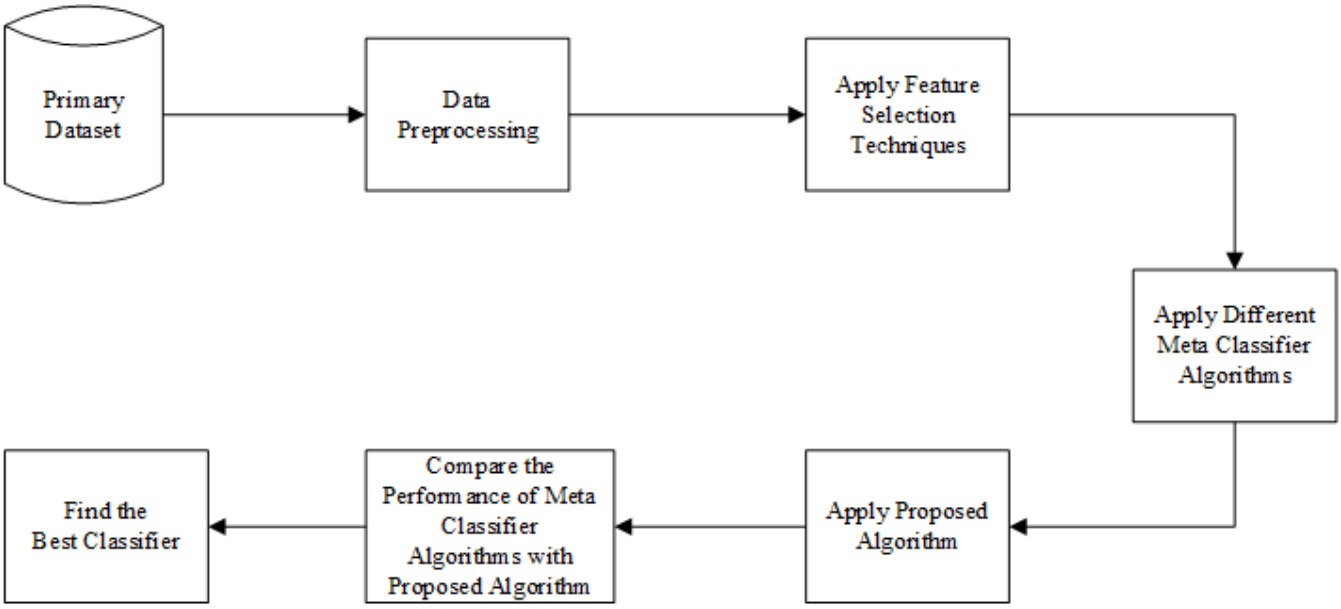

Fig. 1. Proposed Methodology

ated model from the data [1] [2] . Ensemble model improves the performance of classification accuracy of class-imbalanced data.

4. EXPERIMENTAL RESULT

Weka is a software tool which consists of group of machine learning algorithms to accomplish data mining tasks [8]. It includes several tools for data preprocessing, classification, clustering, regression, association rules and visualization. It is also helpful to develop new machine learning methods. In this experiment, we used weka data mining tool to operate features and classification motive. We carried 741 instances of cervical cancer data with 24 attributes. Then we applied many feature selection algorithms such as cfssubsetEvaluator with (AntSearch, CuckooSearch, BeeSearch, BatSearch, ElephantSearch, Evolutionary Search, FlowerSearch, GeneticSearch, FireflySearch, Greedy Stepwise, HarmonySearch, LinearForward Selection, Multiobjective Evolutionary Search, PSO Search, RhinocerosSearch, SubsetSize Forward Selection, WolfSearch). After that we associated 12 meta classifier algorithms using 10 fold cross-validation into our dataset and selected 5 meta classifier algorithms such as Dagging, AdditiveRegression, CVParameterSelection, MultiScheme and MultiSearch. Then we used our new proposed algorithm which is combined with different meta classifiers for their better performance using the result of error rate such as MALE, RMSLE, MAE and RMSE. These performance of error rate can be defined as follows:

—Root Mean Squared Logarithmic Error (RMSLE)
Root Mean Squared Error (RMSE) and Root Mean Squared Logarithmic Error (RMSLE) both are the techniques to find out the difference between the values predicted. RMSLE also measures the ratio between actual and predicted.

$$RMSLE = \frac{1}{N}\sum_{i=1}^{N}(log(p_i+1) - log(a_i+1))^2 \quad (1)$$

—Mean Absolute Error (MAE)

Table 2. Compare Error Rate for Different Classifier.

Classifiers	MALE	RMSLE	MAE	RMSE
Dagging	0.0765	0.1528	0.0955	0.2189
Additive Regrassion	0.0764	0.1564	0.0962	0.2212
CV Parameter Selection	0.0863	0.1593	0.1046	0.2287
Multi Scheme	0.0863	0.1593	0.1046	0.2287
Multi Search	0.0774	0.1552	0.0967	0.2195
Proposed Method	**0.0737**	**0.151**	**0.0929**	**0.2151**

Mean Absolute Error (MAE) is the average vertical distance between each point and the identity line where point i has coordinates (xi, yi). The Mean Absolute Error is given by:

$$MAE = \sum_{i=1}^{N}|y_i - x_i| \quad (2)$$

—Root Mean Squared Error (RMSE)
The root-mean-squared error (RMSE) is a measure of how well the model is performed. It does this by measuring difference between predicted values and the actual values. RMSE is defined as-

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2 \quad (3)$$

Then using these formula, we generated Mean Absolute Logarithmic Error (MALE), Root Mean Square Logarithmic Error (RMSLE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) value for each meta classification algorithms and our proposed algorithm to find out the error rate and compared the result of proposed algorithm with the result of others meta classifier algorithms which is presented in table II. The result of this experiment can help to find the best classier.

Table 2 is considered different evaluation metrics such as MALE, RMSLE, MAE and RMSE of different meta classifier algorithms and proposed algorithm.
In figure 2, the graphical representation of error rate results of different classifiers of cervical cancer dataset is shown. Here it is seen that the percentage of of error rate of proposed algorithm is lower than others meta classifier algorithms.

5. DISCUSSIONS

In this experiment, we explored the cervical cancer data by applying Correlation-based Feature Selection (CFS) methods with many subset classifiers, ranking most significant features with different feature selection method, implementing different meta classifier algorithms and an proposed ensemble learning algorithm into this dataset. After that, we originated different evaluation metrics such as MALE, RMSLE, MAE and RMSE of different meta classifiers which produce the best performing results than others. Then we applied the proposed algorithm with its evaluation metrics such as MALE, RMSLE, MAE and RMSE. Then we compared the error rate of proposed algorithm with error rate of several meta classifiers.

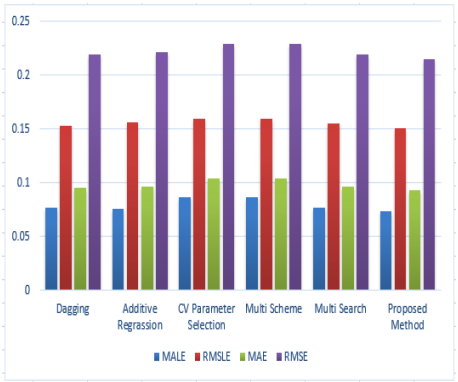

Fig. 2. Error Rate of Individual Classifiers

When we compared the lowest error rate of different meta classifiers with the error rate of proposed algorithm, we found that the proposed algorithm is shown the lowest error for any possible feature sets of cervical cancer dataset. Here we found that all the error rate of MALE (0.0737), RMSLE (0.151), MAE (0.0929) and RMSE (0.2151) of the proposed algorithm is showed lowest rate. Because we know that the classifier which have less error rate evaluated better performance. So we found the better performance for the proposed algorithm. In figure 2, we also shown the compared error result visually in graphical representation which is very easy to compare the result and find the best classifier.

6. CONCLUSION & FUTURE WORKS

In this work, we have used different meta classification techniques on the data set of the factors that cause cervical cancer. Here we have evaluated highest significant ranked features from different feature selection methods and selected five meta classifiers to analyze the dataset and evaluate the lowest error rate for better performance.Then we have applied an ensemble learning proposed algorithm based on AttributeSelected meta classifier which combines with AdditiveRgression classifier. Then we have compared the result of proposed algorithm with others meta classifier algorithms. Finally, we have found that our proposed algorithm gives better performance than the existing meta classifiers. In future, more learning methods will be associated with this proposed ensemble learning method to find out more significant risk factors of cervical cancer which will more acceptable to detect cancer for further research.

7. REFERENCES

[1] Dewan Md Farid, Ann Nowe, and Bernard Manderick. Ensemble of trees for classifying high-dimensional imbalanced genomic data. In *Proceedings of SAI Intelligent Systems Conference*, pages 172–187. Springer, 2016.

[2] Dewan Md Farid, Mohammad Zahidur Rahman, and Chowdhury Mofizur Rahman. An ensemble approach to classifier construction based on bootstrap aggregation. *International Journal of Computer Applications*, 25(5):30–34, 2011.

[3] Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis*, pages 243–250. Springer, 2017.

[4] Sue J Goldie, Louise Kuhn, Lynette Denny, Amy Pollack, and Thomas C Wright. Policy analysis of cervical cancer screening strategies in low-resource settings: clinical benefits and cost-effectiveness. *Jama*, 285(24):3107–3115, 2001.

[5] Seung Hee Ho, Sun Ha Jee, Jong Eun Lee, and Jong Sup Park. Analysis on risk factors for cervical cancer using induction technique. *Expert Systems with Applications*, 27(1):97–105, 2004.

[6] Jorng-Tzong Horng, Kai-Chih Hu, Li-Cheng Wu, Hsien-Da Huang, Feng-Mao Lin, Shir-Ly Huang, Horn-Cheng Lai, and Ton-Yuen Chu. Identifying the combination of genetic factors that determine susceptibility to cervical cancer. *IEEE Transactions on Information Technology in Biomedicine*, 8(1):59–66, 2004.

[7] Xiaoxia Hu, Julie K Schwarz, James S Lewis, Phyllis C Huettner, Janet S Rader, Joseph O Deasy, Perry W Grigsby, and Xiaowei Wang. A microrna expression signature for cervical cancer prognosis. *Cancer research*, 70(4):1441–1448, 2010.

[8] Sushilkumar Rameshpant Kalmegh. Comparative analysis of weka data mining algorithm randomforest, randomtree and ladtree for classification of indigenous news data. *International Journal of Emerging Technology and Advanced Engineering*, 5(1):507–517, 2015.

[9] Nisa M Maruthur, Shari D Bolen, Frederick L Brancati, and Jeanne M Clark. The association of obesity and cervical cancer screening: a systematic review and meta-analysis. *Obesity*, 17(2):375–381, 2009.

[10] Nathalie Reesink-Peters, G Bea A Wisman, Carmen Jéronimo, C Yutaka Tokumaru, Yoram Cohen, Seung Myung Dong, Harrie G Klip, Henk J Buikema, Albert JH Suurmeijer, Harrie Hollema, et al. Detecting cervical cancer by quantitative promoter hypermethylation assay on cervical scrapings: a feasibility study. *Molecular Cancer Research*, 2(5):289–295, 2004.

[11] R Sankaranarayanan, BM Nene, K Dinshaw, R Rajkumar, S Shastri, R Wesley, P Basu, R Sharma, S Thara, A Budukh, et al. Early detection of cervical cancer with visual inspection methods: a summary of completed and on-going studies in india. *salud pública de méxico*, 45(S3):309–407, 2003.

[12] Mark Schiffman, Philip E Castle, Jose Jeronimo, Ana C Rodriguez, and Sholom Wacholder. Human papillomavirus and cervical cancer. *The Lancet*, 370(9590):890–907, 2007.

[13] Kuttiannan Thangavel, P Palanichamy Jaganathan, and PO Easmi. Data mining approach to cervical cancer patients analysis using clustering technique. *Asian Journal of Information Technology*, 5(4):413–417, 2006.

[14] Chih-Jen Tseng, Chi-Jie Lu, Chi-Chang Chang, and Gin-Den Chen. Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Computing and Applications*, 24(6):1311–1316, 2014.