폐암 생존율 이진 분류 성능 분석

벨무루간 아레스 발라지, 최철웅, & 김경백
전남대학교 전자컴퓨터공학부

Analysis of Performance of Binary Classification for Lung Cancer

Velmurugan Arresh Balaji, Chulwoong Choi, & Prof.Kyungbaek Kim
Dept. Electronics and Computer Engineering, Chonnam National University
E-mail: arreshvnass@gmail.com, sentilemon02@gmail.com,
kyungbaekkim@jnu.ac.kr

ABSTRACT

Cancer Patients are feeling better than they have felt in the past. Because of the tremendous changes in the field of oncology, the survival rate of the patients has been increased and the survival period is now longer for many cancers. In this modern era, International Medical research requires Collaborative work with many other departments to predict and analyze the survival period of cancer victims for precise treatments. This paper discusses the analysis and investigation of the best threshold for taking the highest performance for classification of survival period of the Lung cancer patients records obtained from the dataset of the Chonnam National University Hospital, South Korea. Generally, Kaplan-Meier Survival Analysis is the simplest estimation way for computing the survival over time in spite of all these difficulties associated with subjects or situations. But, the maximum lifetime of the records were efficient within 2 years of the given dataset. Hence, we tried to investigate the highest performance month label based on the Decision tree (j48) technique. It is one of the Machine Learning based Binary classification methods. We have obtained the peak ROC of 0.63 for the 10th month. The obtained PR Curve indicates that the F-Measure average and ROC were very low with respect to the Threshold Months (Survival Month), even though it is literally acceptable in the Medical Field.

1. Introduction

Decision tree is one of the most popular machine learning algorithms used all along, Decision trees are used for both classification and regression problems. Decision tress often mimic the human level thinking so it's so simple to understand the data and make some good interpretations. It is a tree where each node represents a feature (attribute), each link(branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value). The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf (or minimize the error in every leaf). For this Survival Period prediction based binary classification, we undergone 24 iterations of Decision tree classification with the help of Weka tool. Weka is a comprehensive software that lets you to pre-process the big data, apply different machine learning algorithms on big data and compare various outputs. This software makes it easy to work with big data and train a machine using machine learning algorithms. And we have classified the dataset based on the Decision tree j48 method for a period of 24 months (2 years).

The two diagnostic tools that help in the interpretation of binary (two-class) classification predictive models are ROC Curves and Precision-Recall curves. The Plots from the curves can be created and used to understand the trade-off in performance for different threshold values when interpreting probabilistic predictions. Each plot can also be summarized with an area under the curve score that can be used to directly compare classification models.

2. Related Works:

As a result, a lot of time is spent searching for the most appropriate machine learning algorithms applicable in clinical prognosis that contains either binary-valued or multi-valued attributes. Multilayer Perceptron, J48, and the Naive Bayes algorithms were used to train and test models on Thoracic Surgery datasets obtained from the University of California Irvine machine learning repository. Stratified 10-fold cross validation was used to evaluate baseline performance accuracy of the classifiers. The comparative analysis shows that multilayer perceptron performed best with classification accuracy of 82.3%, Decision tree J48 came out second with classification accuracy of 81.8%, and Naive Bayes came out the worst with classification accuracy of74.4%. The quality and outcome of the chosen machine learning algorithms depends on the ingenuity of the clinical miner. [1].

3. Decision Tree Based Binary Classification:

In this paper, we have discussed about the obtained ROC Curves and Precision-Recall Curves for imbalanced classification of the CNU Hospital Lung Cancer Dataset. The original Lung cancer CNU Hospital data contains 15 attributes and 673 records. The attributes are Patient_id, Sex, Diagno- -sis_Date, Death_Date, Age_Diagnosis, Mcode, Final_ Stage, T_N_M_Stage, Final_Stage_Date, Smk_Exp, Amt_Smk_Per_Day, PeriodSmk, GabNyunSmk, Smk_ Stop Year and AgeAtDeath Respectively. For the Decision tree based Binary Classification we pre-processed the Dataset, by adding a new column called Survival_Days. Here, Survival_Days = (AgeAt _Death ~ Age_Diagnosis).

4. Evaluation:

we had tested this dataset through Decision tree algorithm with the help of Weka v3.9.3 a Machine Learning based Data mining tool for the binary classification techniques. For the testing we used the pre-processed dataset of 9 columns and 673 rows. The attributes are Age_Year, Smk_YN, Smk_Daily, Smk_Period Year, Smk_Year, T_Stage, N_Stage, M_Stage and the Different_Survival_ Label. Here the Differernt_Survival_Label ranges from 1 to 24 labels(attributes) for the 673 patient records. Each label (X) have the constraint as IF (Survival_Days > X *30) the Label value is "Live" else "Dead". Where, X=1,2,3 ,…,24.

By running these 24 iterations on Decision tree (j48) algorithm, we can get the ROC and F Measures of 24 months respectively. F Measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. For efficient results, F1 measure should be above 50% (0.5). With The F-Measures and ROC values on Y-axis and Threshold Months on the X-axis. we can plot the PR Curve and ROC curve for the analysis.

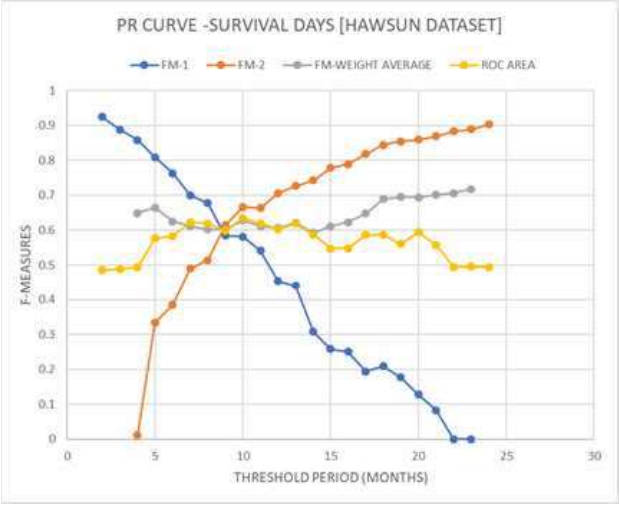

Figure 1 : Result Graph PR AND ROC CURVE (F-Measures Vs Threshold Months)

In the Figure 1, The PR curve and ROC curve were plotted. It clearly shows that, the peak Regions of Convergence (ROC) was obtained at the threshold month 10 at 0.63 F measure. And FM-1 is the F measure of the Live patients and FM-2 is the F measure of the Death patients. There is a threshold point at month 10 where each parameter collide each other.

5. Conclusion

From the Obtained results, the Binary classification done was good upto 10 month Labels for the given CNU Hospital Dataset. According to Precision Recall Curve the minimum F measure is 0.5. In our research, the maximum peak obtained is only 0.63, which is considered to be a very poor result, even though it is literally acceptable in the Medical Field. In order to improve the results, the CNU Hospital have to provide some huge volume of the dataset, because the tested dataset is not sufficient and suitable for this Decision tree J48 based Binary Classification Approach.

Acknowledgements

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961).

References

[1] Danjuma, Kwetishe Joro. VBVet al. “Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the lung cancer Patients." ArXiv abs/1504.04646 (2015):n.pag.

[2] Friedl, Mark A. and Carla E. Brodley, et al. "Decision tree Classification of land cover from Remotely Sensed Data." Remote Sensing of Environment, Volume61, Issue 3 (1997), ISSN 0034-4257, doi:10.1016/s0034-4257(97)00040-7.

[3] Mohammad M. Ghiasi, Sohrab Zendehboudi, Ali Asghar Mohsenipour, Decision tree-based diagnosis of coronary artery disease: CART model, Computer Methods and Programs in Biomedicine, Volume192, 2020, 105400, ISSN 0169-2607, doi.org/10.1016/j.cmpb.2020.105400.

[4]Shankru Guggari, Vijayakumar Kadappa, V. Umadevi, Ajith Abraham, Music rhythm tree based partitioning approach to decision tree classifier, Journal of King Saud University - Computer and Information Sciences, 2020, ISSN 1319-1578, doi.org/10.1016/j.jksuci.2020.03.015.

[5] Mohmad Badr Al Snousy, Hesham Mohamed El-Deeb, Khaled Badran, Ibrahim Ali Al Khlil, Suite of decision tree-based classification algorithms on cancer gene expression data, Egyptian Informatics Journal, Volume12, Issue 2, 2011, Pages73-82, ISSN 1110-8665, doi.org/10.1016/j.eij.2011.04.003.