SURVIVAL PREDICTION OF LUNG-CANCER PATIENTS AFTER ONE YEAR OF THORACIC SURGERY

PROJECT DESCRIPTION

The World Health Organization (WHO) reports cancer as the leading cause of death worldwide. Moreover, the organization lists lung cancer on top of the list of all cancer types in terms of the number of cases and deaths globally. On the other hand, the Centers for Disease Control and Prevention (CDC) list lung cancer as the third most common cancer in the United States. Besides, more people die from lung cancer in the US than any other type of cancer, equally among men and women according to CDC.

Medical records, laboratory, and examination data, as well as questionnaires are continually conducted worldwide to assess and quantify the risk factors and reduce the impact of this global problem through early detection and early based prevention programs. Examining patterns in clinical analysis and finding correlations between symptoms and risk factors can help reduce the implications of lung cancer by avoiding the causes far before the detection of the disease. More specifically, this project is based on medical data that is related to post-operative life expectancy in lung-cancer patients. Such dataset lists the features that influence the survival or mortality of patients one year after undergoing thoracic surgery.

In this project, we will use medical examination data that report mortality rates of lung cancer patients and build machine learning predictive models to predict the survival/mortality of patients one year after the thoracic surgery. The approach is to devise comparable machine learning algorithms and implement them in this medical application to predict post-operative life expectancy in lung-cancer patients. The topic described is a supervised learning task, more specifically, a classification problem with risk of mortality as a target variable. The target describes the risk of thoracic surgery on patients after a one-year timespan.

Part of the goals of this project is to also identify the factors that contribute to increased risk of mortality among patients and to determine the most predictive features of the dataset that are sufficient to predict the patients' survival.

TEAM MEMBERS AND ROLES

Both team members will contribute to advance the development of this project. However, the primary tasks assigned to each team member are as follows:

Manal Zneit – will implement the machine learning algorithms and build predictive models.

Sabina Bhuiyan – will design the visualization tool that will demo the comparisons of the models.

Both team members will implement evaluation techniques for model selection.

DATASET

The project is based on a dataset collected from patients at Wroclaw Thoracic Surgery Centre in Poland where the data was collected retrospectively during the years 2007 and 2011. The dataset is listed in the UCI repository in the Life Sciences category under the name "Thoracic Surgery Data" and it has a total of 470 instances and 16 attributes and 1 target vector. The values of the attributes are mostly categorical, and the target is a binary-valued vector representing survival or death of each patient instance.

The dataset is challenging since it is imbalanced and has an unequal class distribution (70 True/ 400 False instances). Traditional training algorithms when applied on imbalanced datasets will tend to make decisions biased toward the majority class [1]. Previous work proposed a solution approach to solve imbalanced data problems on this dataset where an approach to extract decision rules to solve imbalanced data problems is presented.

CSCI 353/795 RELATED TOPICS

The project described is a real-world problem – mainly, a classification task where we will implement several machine learning classifiers that predict the survival of patients who underwent the surgery after a one-year period. SVM, artificial neural network (perceptron), an instance-based learning model (K-Nearest Neighbors), and a probabilistic classifier (Naïve Bayes) will be implemented and compared accordingly. These models will be compared and evaluated using binary classifiers evaluation techniques. We will perform cross validation and choose optimal hyperparameters for model selection. In addition, since this project builds medical predictive models, evaluation metrics such as sensitivity and specificity will be used primarily, as well as other metrics such as accuracy, F1-score, ROC curve, and AUC. The results will be compared with the results of publications that applied comparative ML algorithms on this dataset.

TIMELINE

Week 1: Algorithms development and coding (SVM and Perceptron).

Week 2: Algorithms development and coding (KNN and Naïve Bayes).

Week 3: Evaluation techniques and code improvement.

Week 4: Draft demo and report.

Week 5: Finalize code.

Week 6: Finalize presentation, report, and visualization tool.

REFERENCES

[1] Zieba, M., Tomczak, J., Lubicz, J., & Swiatek, J., Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, vol. 14 (2013), 99-108. DOI: 10.1016/j.asoc.2013.07.016

[2] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* 20, 16, (2020). DOI: 10.1186/s12911-020-1023-5

[3] Desuky, A. S., Bakrawy, L. M. E. Improved Prediction of Post-operative Life Expectancy after Thoracic Surgery. *Advances in Systems Science and Applications*, 16(2), 70-80, (2016).

[4] Nachev, A. and Reapy, T. Predictive models for post-operative life expectancy after thoracic surgery. *Mathematical and Software Engineering*, 1(1), 1-5, (2015).