ISSN: 2508-7894 © 2019 KAIA. http://www.kjai.or.kr
Doi: http://dx.doi.org/10.24225/kjai.2019.7.2.19

A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning

[1] Yu-Jin NAM, [2] Won-Ji SHIN

[1, First Author] Student, Dept. of Bio Medical Engineering, Eulji University, Korea, Email: skadb4513@gmail.com
[2, *Corresponding Author] Dept. of Radiological Technology, Eulji University, Korea, Tel: +82-31-740-7361, Email: wonji0423@gmail.com

Received: November 30, 2019., Revised: December 04, 2019., Accepted: December 10, 2019.

Abstract

Lung cancer is a chronic disease which ranks fourth in cancer incidence with 11 percent of the total cancer incidence in Korea. To deal with such issues, there is an active study on the usefulness and utilization of the Clinical Decision Support System (CDSS) which utilizes machine learning. Thus, this study reviews existing studies on artificial intelligence technology that can be used in determining the lung cancer, and conducted a study on the applicability of machine learning in determination of the lung cancer by comparison and analysis using Azure ML provided by Microsoft. The results of this study show different predictions yielded by three algorithms: Support Vector Machine (SVM), Two-Class Support Decision Jungle and Multiclass Decision Jungle. This study has its limitations in the size of the Big data used in Machine Learning. Although the data provided by Kaggle is the most suitable one for this study, it is assumed that there is a limit in learning the data fully due to the lack of absolute figures. Therefore, it is claimed that if the agency's cooperation in the subsequent research is used to compare and analyze various kinds of algorithms other than those used in this study, a more accurate screening machine for lung cancer could be created.

Keywords: Lung Cancer, Machine Learning, Two-Class Support Vector Machine, Two-Class Decision Jungle, Multi-Class Decision Jungle

1. Introduction

Lung cancer is a chronic disease which ranks fourth among cancer incidences. It takes 11 percent of the total cancer incidences in South Korea. In addition, the long-term survival rate of lung cancer patients in the terminal stage is less than 20 percent, due to easy metastasis and poor prognosis. Early detection and treatment of the lung cancer is the most effective way to reduce the lung cancer mortality (Aberle, 2011). In order to solve such problem, there is an active study on the usefulness and utilization of Clinical Decision Support System (CDSS) using machine learning (ML). However, the application of these artificial intelligence technologies is still in the beginning stage (Lee, 2017). Thus, this study reviewed studies regarding artificial intelligence technology that can be used to determine lung cancer, and conducted a study on the applicability of machine learning in determination of the lung cancer by comparing and analyzing the accuracy of Azure's algorithms provided by Microsoft. Specifically, we utilized Two-class Support Vector Machine, Two-class Boosted Decision Tree, Two-class Decision Tree, Two-class Logistic Regression Algorithms, and the Lung Cancer data provided by Kaggle is utilized to create a discriminative model.

2. Literature review

2.1 Microsoft Azure and Machine Learning

Microsoft Azure Machine Learning includes cloud services that enable the creation, deployment and management of applications by developers through a global network of data centers. This cloud computing model emphasizes the cloud platform's differentiating features namely flexibility and scalability.

Azure ML also supports multiple ML algorithms related to regression, classification, and clustering. It allows customization of the models using python and R (Qasem et al., 2015). Azure ML studio allows dragging and dropping of modules and datasets (i.e., Ml algorithms, feature selection and pre-processing) and links them together. It supports about 100 techniques that address ML algorithm, function selection and data preprocessing, navigation, modeling result verification and method selection, regression, classification, text analysis, and recommendations.

Machine learning tasks are typically classified into three broad categories: 1) supervised learning, in which the system infers a function from labeled training data, 2) unsupervised learning, in which the learning system tries to infer the structure of unlabeled data and 3) reinforcement learning, in which the system interacts with a dynamic environment (Hansan, 2014).

2.2 Machine Learning Algorithms

2.2.1. Two-Class Support Vector Machine

Support vector machines (SVMs) are well-researched instance of the supervised learning. This implementation is suitable to prediction of two possible outcomes, based on either continuous or categorical variables.

2.2.2. Two-Class Decision Jungle

Decision jungle is a recent extension of the decision forests. A decision jungle consists of an ensemble of decision directed acyclic graphs (DAGs). By allowing tree branches to be merged, a decision DAG typically has a lower memory footprint and better generalization performance than a decision tree, albeit at the cost of somewhat longer training time. Decision jungles are non-parametric models that can represent non-linear decision boundaries. They perform integrated feature selection and classification and are resilient in the presence of noisy features. In this case, Two-Class Decision Jungle contains instances of only 2 classifications (class 0 and 1).

2.2.3. Multi-Class Decision Jungle

Basically, Multi-Class Decision Jungle has the same method that Two-Class Decision Jungle does. The difference is that it contains at least two instances of classification (from class 0 to class n).

3. Methodology

Our methodology consists of three main steps; the first step is data set selection. The second step includes preprocessing in which the original data is prepared for classification. The last step contains training models according to the accuracy of each algorithm.

3.1. Dataset

Publicly available dataset has been utilized which was obtained from Kaggle from this research (Talha, 2019). The dataset contains 59 patient cases and 7 attributes which include the patient name, surname, age, smokes, areaQ, alkhol and result.

Table 1. A part of Dataset

Name	Surname	Age	Smokes	AreaQ	Alkhol	Result
John	Wick	35	3	5	4	1
John	Constantine	27	20	2	5	1
Camela	Anderson	30	0	5	2	0
Alex	Telles	28	0	8	1	0
Diego	Maradona	68	4	5	6	1
Cristiano	Ronaldo	34	0	10	0	0
Mihail	Tal	58	15	10	0	0
Kathy	Bates	22	12	5	2	0
Nicole	Kidman	45	2	6	0	0
Ray	Milland	52	18	4	5	1
Fredric	March	33	4	8	0	0
Yul	Brynner	18	10	6	3	0
Joan	Crawford	25	2	5	1	0
Jane	Wyman	28	20	2	8	1
Anna	Magnani	34	25	4	8	1
Katharine	Hepburn	39	18	8	1	0
Katharine	Hepburn	42	22	3	5	1
Barbra	Streisand	19	12	8	0	0
Maggie	Smith	62	5	4	3	1

3.2 Data Preprocessing

Machine Learning fundamentally depends on the quality of data. Raw data tend to be vulnerable to noise, missing values, outliers and inconsistency. So, it is vital for selected data to be processed before being learned. Preprocessing the data is an essential step to enhance data efficiency. Data preprocessing is one of the most important data processing steps, which handles data preparation and transformation of data sets, making knowledge discovery more efficient. There are following steps which were used to preprocess data in this study for the experiments.

Step 1: Select columns in dataset. This module is particularly useful when there is a need to limit the columns available for a downstream operation, or when there is a need to reduce the size of the dataset by removing unneeded columns. The columns in the dataset are output in the same order as in the original data, even if they are specified in a different order. To choose necessary columns for experiments, attributes of Age, Smokes, AreaQ, Alkhol and Result are selected.

Step 2: Edit metadata. There are two reasons to use editing metadata: arrange in the order desired and rename columns to make it easier to examine. First, set the dataset to integer type. Then, rename the dataset from Indonesian to English.

Step 3: Clean missing data. The goal of such cleaning operation is to prevent problems caused by missing data that can arise during training of a model. In this research, this module is used to replace missing values with a placeholder for mean or other value.

Step 4: Filter-based feature selection. In general, *feature selection* refers to the process of applying statistical tests to inputs when a specified output is given, to determine which columns in the output are more predictive. This module makes columns with poor feature selection scores disappear in the dataset.

3.3 Training Models according to algorithms

The data was pre-processed by following the steps mentioned in section 3.2 (Select columns in Dataset, Edit Metadata, Clean Missing Data, Filter Based Feature Selection) and it is divided into training data and test data by the ratio of 7:3 using Split Data. Training data was used in the purpose of training each algorithm, and test data is used to evaluate the trained algorithms. Overall, Two-Class Support Vector Machine algorithm and the Multi-Class Decision Jungle algorithm are compared. And Two-Class Decision Jungle algorithm is analyzed using the Evaluate Model.

4. Results

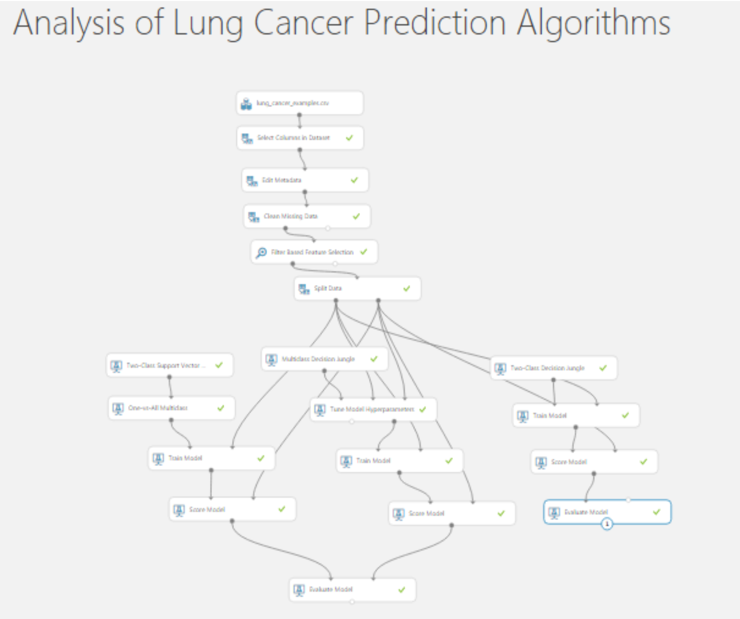
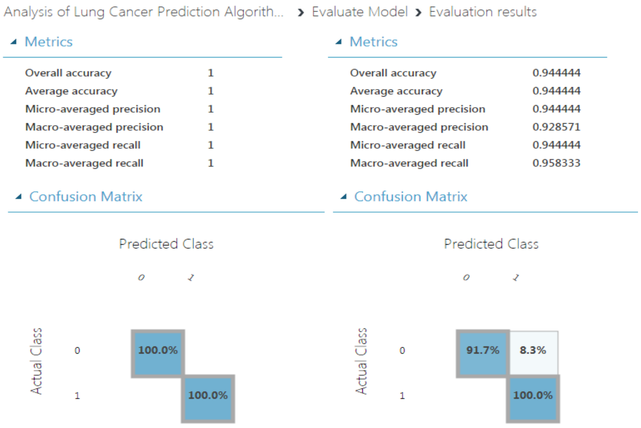

Analysis of Lung Cancer Prediction Algorithms
lung_cancer_examples.csv
Select Columns in Dataset
Edit Metadata
Clean Missing Data
Filter Based Feature Selection
Split Data
Two-Class Support Vector ...
Multiclass Decision Jungle
Two-Class Decision Jungle
One-vs-All Multiclass
Tune Model Hyperparameters
Train Model
Train Model
Train Model
Score Model
Score Model
Score Model
Evaluate Model
Evaluate Model

Figure 1. Analysis and Comparison of the Lung Cancer Prediction in Azure ML studio

The model in this study was built in Azure ML using three different algorithms: Two-Class Decision Jungle, Multi-Class Decision Jungle and Two-Class Support Vector Machine. Figure 1 shows the final workflow of the algorithm. To look at the overall workflow, we loaded 'Lung Cancer_example.csv' and selected the required columns. After that, the name and data type were specified, and the filter was outputted with the optimized properties. In order to split the training data and the test data, the ratio of the split data was set to 7: 3 and based on the classifications, two-class support vector machine, the two-class support vector machine, and the multi-class decision jungle were classified. Training data was used for each of three algorithms, and test data was trained by putting values based on the split data divided above.

The evaluation step measured the trained model's accuracy. Scored dataset and scored trained dataset were the inputs to this step (Osama, 2018). As a result, two evaluate models were visualized. Metrics and Confusion Matrix are represented for each of Support Vector Machine and Multi Class Decision Jungle, and ROC curve represents the result curve of Two Class Decision Jungle. In Figure 2, Metrics represents the sources of Azure monitor metrics and Confusion Matrix is used to describe the performance of a classification model.



Analysis of Lung Cancer Prediction Algorith... › Evaluate Model › Evaluation results

Metrics

Overall accuracy	1
Average accuracy	1
Micro-averaged precision	1
Macro-averaged precision	1
Micro-averaged recall	1
Macro-averaged recall	1

Confusion Matrix

Actual Class \ Predicted Class	0	1
0	100.0%	
1		100.0%

Metrics

Overall accuracy	0.944444
Average accuracy	0.944444
Micro-averaged precision	0.944444
Macro-averaged precision	0.928571
Micro-averaged recall	0.944444
Macro-averaged recall	0.958333

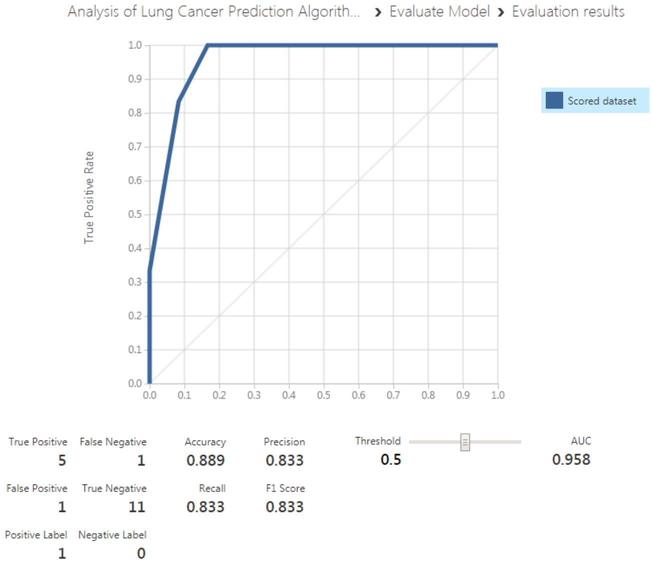
Confusion Matrix

Actual Class \ Predicted Class	0	1
0	91.7%	8.3%
1		100.0%

Figure 2. Result of Two-Class Support Vector Machine and Multi-Class Decision Jungle

The metrics on the left are Two-Class Support Vector Machine algorithm. The overall accuracy was 100%, the highest among the three algorithms. In addition, the prediction according to the prediction class and the actual class in the confusion matrix shows that the ratio of true positive and false negative is 100%, respectively.

The metrics on the right are Multi Class Decision Jungle algorithms, with the second highest overall accuracy in prediction. The figure is about 94%. In the confusion matrix, the ratio of true positive is 91.7% and that of false negative is 100%. Compared to the two class Support Vector Machine algorithm, the true negative ratio was about 8.3%.



Analysis of Lung Cancer Prediction Algorith... › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
5	1	0.889	0.833	0.5	0.958
False Positive	**True Negative**	**Recall**	**F1 Score**		
1	11	0.833	0.833		
Positive Label	**Negative Label**				
1	0				

Figure 3. Result of Two-Class Decision Jungle (ROC curve)

The Two-Class Decision Jungle algorithm shows a ROC curve with an accuracy of 88.9%. It can be seen that two-class decision jungle method has the lowest predictive value compared to the accuracy the other two algorithms as the confusion matrix presents above. The diagnosis rate of the lung cancer varies from one test to another, with the averaging of 60%, diagnosis rate in bronchial lung biopsy, while that of cytogenetic assay averaging 80%. Thus, it can be concluded that the algorithm analysis used in this study has higher accuracy in diagnosis.

5. Conclusion

In this study, the recent literature was reviewed with visualization of machine learning methods in Azure ML studio (Kavakiotis, 2017). Three algorithm's techniques were used as a result of this study in which all of them produce different results. The Support Vector Machine algorithm showed the most ideal accuracy. However, it was judged that the result was due to the errors due to limitations in the number of data, which resulted in the difficulty in the application of it to the determination of the lung cancer. In particular, Multi-Class Decision Jungle algorithm and Two-Class Decision Jungle algorithm showed 5.5% deviation. Among them, the most accurate technique is the Multi-Class Decision Jungle algorithm. Compared to two-class algorithm, it is believed that it classified data in various ways. As a result, Azure ML can also produce visible results in comparing and analyzing algorithms, which can help users understand intuitively.

This study has limitations in the size of the 'Big Data' used in Machine Learning. Although the data provided by Kaggle, which was used in this study, is most suitable, it is assumed that there is a limit to learn the data fully due to the lack of absolute figures. In addition to Kaggle, we requested the big data on lung cancer to the Health Insurance Review & Assessment Service Institute, the National Cancer Center, and the Korean Association for Lung Cancer, but it was impossible to obtain big data on lung cancer that was suitable for domestic conditions without the cooperation of the agencies, since the data was not disclosed to the public during the project. Therefore, it is judged that if there would be agency's cooperation in the subsequent research for comparing and analyzing various kinds of algorithms other than those used in this study, a more accurate screening machine for the lung cancer could be created.

References

Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., & Clapp, J.D. (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine, 365,* 395–409.

Alam, T. M. (2019). Cervical Cancer Prediction through Different Screening Methods using Data Mining (IJACSA). *International Journal of Advanced Computer Science and Applications*, *10*(2),388-396

Harfoushi, O. (2018). Sentiment Analysis Algorithms through Azure Machine Learning: Analysis and Comparison. *Modern Applied Science*, *12*(7), 49-58.

Ioannis Kavakiotis, OgaTsave, Athanasios Salifoglou, Nicos, Maglaveras, Ioannis, et al. (2017) Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal, 15*, 104-116.

Qasem, M., Thulasiram, R. K., & Thulasiram, P. (2015). Twitter sentiment classification using machine learning techniques for stock markets, *IEEE, 10*, 834-840.

Lee, J.G., Jun S., Cho, Y.W., Lee, H., Kim, G.B., Seo, J.B., & Kim, N. (2017). Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, *18*(4), 570-584.

Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd Ed.). New Jersey, USA: Prentice Hall.