(Vajratiya Vajrobol)

The research synopsis

The classification of air quality – lung cancer data using machine learning technique

Introduction

Lung cancer is a type of cancer that begins in the lungs and this kind of cancer leads to death worldwide. The majority of people who smoke are likely to get lung cancer and some who never smoked also have a chance to get lung cancer.

According to WHO, air quality is also classified as a human carcinogen similar to tobacco and smoke. Air pollution is a leading cause of cancer death worldwide. Several previous studies have found that the high composition of particulate air pollution with high death rates.

The proper understanding of lung cancer related to air quality is necessary to interpret the underlying cause and risk factor for cancer. Combining the lung cancer incident rates available from National Cancer Institute and advanced machine learning techniques may bring out the strongest factor that has more accurate prognostic value. In this paper, we will analyze the air quality factors associated with lung cancer.

Literature review

Lung cancer is a type of cancer that begins in the lungs and this kind of cancer leads to death worldwide. The majority of people who smoke are likely to get lung cancer and some who never smoked also. Many researchers have worked on predicting lung cancer with different machine learning algorithms and data mining techniques, such as rule-based system decision trees, Naïve Bayes, and Artificial neural network. Krishnaiah et al. [4] present an effective model to predict patients with lung cancer disease appeared to naïve-Bayes followed by If-then rule, Decision tree, and Neural network.

Furthermore, Rajan et al. [7] predict lung cancer at an early stage with Artificial Intelligent techniques and Data mining and the research can increase survival rates by five years. A study conducted by Lakshmanaprabu et al.[5] focused on applying image processing and machine learning method such as deep learning to predict the CT scan picture. The result found that the accuracy, sensitivity, and specificity with its values 94.56 % and 96.2% and 94.2 respectively which is deeply proficient to classify benign and malignant lung cancer.

Apart from identifying lung cancer patient, there are also different aspects of research; for example, identifying the factors that effects on lung cancer hospitalization expense using conducted by C5.0 algorithm. Yu et al.[8] Identified that the combination of different factors such as gender, diagnosis, main therapy, condition with discharge, the stage of cancer, age, ICU time, admission time, medical fee, laboratory fee, examination fee, surgery fee, nursing fee, and bed charge to predict the medical fee for lung cancer patient and the highest accuracy of model reaches 84 %.

Focusing on air quality and lung cancer, there are several studies conducted on whether air pollution contributes to the occurrence of lung cancer. Cohen et al. [1] found that the air quality is associated with excess lung cancer risks. In addition, Fajersztajn et al. [2] mentioned that the high level of carbon monoxide is an accepted cause of illness but the evidence of nitrogen dioxide is less certain. The most common cause of lung cancer is secondhand smoke. Katsouyanni et al.[3] from many countries have shown elevated risks of lung cancer in urban or industrially polluted areas, generally by up to 1.5 times, even when adjustment for smoking has been attempted. Nyberg et al.[6] observed that there is the association of lung cancer and urban air pollution by using geographical information system (GIS) techniques to assign individual exposures to ambient air pollution from oxides of nitrogen (NOx), nitrogen dioxide (NO2), and sulfur dioxide (SO2) from defined emission sources.

Aims & Objectives

The research aims to discover the knowledge from air quality –lung cancer dataset from Harvard repository and produce the reasonable pattern and relationship that effects on lung cancer. The success of this project will be measured by using accuracy which is, this term tells us how many right classifications were made out of all the classifications. Precision presents out of all that were marked as positive, how many are truly positive. Furthermore, Recall is the actual real positive cases, how many were identified as positive. Also, the F1-score . F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test

Research Methodology

The purposed machine learning algorithm handled air quality-lung cancer data are k-nearest neighbors, discriminant analysis, and classification tree. This research tried to investigate the air quality factor that affects lung cancer. The programming language that will be implementing is python. Python is the most popular language for Artificial Intelligence and Machine Learning with a lot of libraries such as scikit-learn, pandas, Keras, TensorFlow, and Matplotlib.

The studies will go through two phases; data processing and applying classifiers. The data processing intends to prepare a dataset for the second phase. The second phase includes using classifiers mentioned before to construct a higher accurate prediction model for air quality – lung cancer data. (https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/HMOEJO/D3GVNB&version=1.0)

A. Data Preprocessing

Even if a dataset is collected in the right format, it may still need processing to apply a machine learning technique and increase the quality of the analysis result. There are a lot of preprocessing techniques. In this study, the method used is data cleaning, dimension reduction, standardization, transformation, and discretization.

Some attributes of air quality-lung cancer data contain missing data. However, these columns will not be used in the data analysis phase. Furthermore, outliers need to be removed from the dataset. Dimension reduction is preceded by selecting attributes based on information gain for the class. Only the significant contributing factors will be selected.

B. Data Analysis

The different classifiers such as K-nearest neighbor, discriminant analysis, and classification tree will be used. Firstly the data is divided into training data and the test data set. The final decision is based on the majority vote of the class. Applying voting to classification algorithms is showing successful improvement in the accuracy of these classifiers.

C.K-nearest neighbor

K-nearest neighbor classifier is a popular method in a wide range in classification problems due to its simplicity and relatively high convergence speed. KNN considers the k nearest instances {il, i2,.., ik} from an instance (x) and decides upon the most frequent class in the set {c1, c2, ... ck}. The most frequent class is assumed to be the class of that instance (x). In order to determine the nearest instance, KNN technique adopts a distance metric that measures the proximity of instance x to k of stored instances. Various distance metrics can be used, including the Euclidean which is used in this paper because it performs well when the continuous attributes are normalized so that they have the same influence on the distance measure between instances. Furthermore, the data dimensionality has been reduced to prevent or reduce its affecting on the performance of the Euclidean distance.

However, there are many disadvantages of KNN classifiers. The main one is the large memory requirement needed to store the whole training set. If the training set is large, response time will be also large which resulted in poor run-time performance. Despite the memory requirement, KNN in general has a good performance in classification problems. Moreover, KNN is very sensitive to irrelevant or redundant attributes and thus affect on the classification accuracy. Hence, the selected dataset should be prepossessed with careful attribute selection technique. Another disadvantage of KNN is the selection of k. If k is too small, then the result can be sensitive to noise. If k is too big, then the result can be incorrect where neighbors include too many points from other classes.

D. Linear Discriminant analysis

Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-reparability in order to avoid over fitting and also reduce computational costs.

E. Classification tree

Classification Trees is a nonparametric method which uses recursive binary partitioning to create a binary tree. The CART algorithm is as follows

1) Initialize the tree containing the training data

2) Obtain a set of binary splits based on one variable

3) Select the best split at a node by estimating impurity functions, the Gini index or entropy

4) Obtain the right-sized tree using Independent test set, or 10-fold cross-validation, or 1-SE rule 5) Assign every terminal node to a class

5) Assign every terminal node to a class

References

[1] Cohen, A. J., & Pope 3rd, C. A. (1995). Lung cancer and air pollution. *Environmental health perspectives*, *103*(suppl 8), 219-224.

[2] Fajersztajn, L., Veras, M., Barrozo, L. V., & Saldiva, P. (2013). Air pollution: a potentially modifiable risk factor for lung cancer. *Nature Reviews Cancer*, *13*(9), 674-678.

[3] Katsouyanni, K., & Pershagen, G. (1997). Ambient air pollution exposure and cancer. *Cancer causes & control*, *8*(3), 284-291.

[4] Krishnaiah, V., Narsimha, G., & Chandra, D. N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*, *4*(1), 39-45.

[5] Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. *Future Generation Computer Systems*, *92*, 374-382.

[6] Nyberg, F., Gustavsson, P., Järup, L., Bellander, T., Berglind, N., Jakobsson, R., & Pershagen, G. (2000). Urban air pollution and lung cancer in Stockholm. *Epidemiology*, 487-495.

[7] Rajan, J. R., & Prakash, J. J. (2013). Early diagnosis of lung cancer using a mining tool. In *National Conference on Architecture, Software systems and Green computing-2013 (NCASG2013)*.

[8] Yu, T., He, Z., Zhou, Q., Ma, J., & Wei, L. (2015). Analysis of the factors influencing lung cancer hospitalization expenses using data mining. *Thoracic cancer*, *6*(3), 338-345.

Conclusion

The application of K-nearest neighbor, discriminant analysis, and classification tree algorithm to determine the relationship between the air quality factors and lung cancer. If the data pattern is extracted, it can reduce lung cancer occurrence rate by avoiding specific air pollution such as PM2.5, Toxic gas components like Carbon monoxide. Moreover, the number of the patient will be reduced when we found the significant factors that need to be avoided. Furthermore, the city identified in the dataset can use the insight for environment and gas emission plans. The research will contribute to increasing the livable level of the city.

This study has some limitations because variables such as the data include only a few years and specify only some part of the United States area. Further studying can be conducted in several places in the world to discover the pattern of data. The dataset does not combine individual information levels but using overall geographical collected data.
