IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Application of Knowledge Engineering for Prediction of Lung Cancer

[1*]N.Vijayalakshmi, [2] J.PolleyAmilya

[1,2]Dept. of Computer Science, Shrimati Indira Gandhi College, Bharathidasan University, Trichy, Tamilnadu, India

Corresponding Author: nvijimca@gmail.com, Tel: +91 9965779358

Available online at: www.ijcseonline.org

Accepted: 18/July/2018, Published: 31/Jul/2018

Abstract: People suffering from Lung cancer are commonly found throughout the world. There are many people who die of this fatal ailment every year. Even though there are many reasons for the occurrence of this disease, it is difficult to say if a person is suffering from lung cancer, unless we test for it. This is costly. It would be highly useful if we could identify significant symptoms in a person that could help to predict the possibility of occurrence of this disease in them. With this objective in mind, we have used a dataset consisting of values of 16 significant symptoms in people who were diagnosed and tested for lung cancer. Based on knowledge of various factors leading to lung cancer and the general symptoms in people suffering from lung cancer, we have chosen these biomarkers. Statistical techniques were used to study and analyze the occurrence of each symptom / factor in the sample population with and without lung cancer, to arrive at a set of predominant symptoms in people with lung cancer. The dataset was also tested for significance of factors using a data mining tool. The dataset was classified using various decision tree algorithms and the results were compared. Decision rules were also generated using Apriori Analysis. This was used to build three data models that help to predict the occurrence of lung cancer among patients with significant symptoms. A maximum accuracy of prediction of 93% was achieved. This was found to be better than prediction through decision tree classifiers.

Keywords: Data Mining, Classification, Prediction, Lung Cancer, Apriori Analysis

I. INTRODUCTION

This research work uses data mining processes and tools to mine new knowledge from record sets of patients suffering from lung cancer. Several factors found in patients with lung cancer have been taken for the study and significant factors out of all factors were identified through statistical analysis. Classification of the given knowledge base using the significant factors is carried out using various decision tree algorithms like Naïve Bayes, Random trees, Random forest and J48 techniques. Accuracy of classification is also compared. Important decision rules obtained through Apriori analysis is used to build a model for prediction of lung cancer among new patients using their values for the given significant factors. Weka has been used for classification and MS-Excel for statistical analysis and prediction. Thus data mining has been used to derive new knowledge pertaining to occurrence of lung cancer among patients with significant symptoms. This is used to automate the prediction of lung cancer among patients based on their medical examination and results.

The rest of the paper is organized as follows: Section I contains the introduction to the research work. Section II contains related work in the area of data mining and lung cancer. Section III contains the methodology used in this research work. Section IV contains results obtained and Section V discusses them in detail. Section VI contains concluding remarks.

II. RELATED WORK

In a study conducted by M.Venkat Dass et al., biomarkers are used to predict the type of lung cancer in patients to aid in the choice of correct cancer treatment strategies. J48 classification decision tree induction algorithm improved by cross validation techniques is used to accurately predict the cancer type. [1] In another study conducted by Juliet Rani Rajan and Chilambu Chelvan, data mining techniques have been used to predict the occurrence of lung cancer in patients at stage 1 using symptoms. [2] A study by Ahmed et al. uses K-means clustering for identifying relevant and non-relevant data from a dataset of 400 records of patients with and without cancer. Significant frequent patterns are discovered using Apriori Tid and decision trees. These are used to predict lung cancer in patients.[3] Ada and Rajneet Kaur in a study of digital X-rays of patients suspected to suffer from lung cancer have used SVM and neural networks to classify the type of lung cancer in patients with abnormal X-rays. They use feature selection to classify X-

rays into normal and abnormal. [4] In yet another study by Yang et al., the researchers have used data mining for cancer staging diagnosis. They try to correlate between pathology report and clinical information. They generate interesting rules and evaluate them [5] Dr.Thangaraju in his study using 300 X-ray chest films, uses feature extraction to classify them as normal and abnormal. He also uses neural networks for classification and image mining. [6] Kawsar Ahmed et al., use K-means clustering for identifying relevant and irrelevant data to lung cancer. Significant frequent patterns are discovered using Apriori Tid and Decision tree algorithms. Using these techniques the team has developed a prediction system that provides a person's predisposition to lung cancer. [7]

In a doctoral study by Guoxin Tang, data mining was used to analyze and diagnose lung cancer. The research tries to examine the relationship between patient outcomes and conditions of the patients undergoing different treatments for lung cancer and to develop models to predict the mortality of lung cancer. The study identifies demographic, finance and clinical factors related to the diagnosis or mortality of lung cancer to help physicians and patients in their decision making. Text mining and cluster analysis, decision tree classification, times series and predictive modeling were used in the study. Significant factors for the mortality of lung cancer were also identified. [8] In a study by V.Krishniah et al., a prototype lung cancer prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer database. It uses Naïve Bayes classification and neural networks to classify and predict the disease.[9] Suresh H.Moolgavkar et al. use patterns of exposure to radon and cigarette smoke experienced by individuals to study the proliferation of lung cancer. Age specific relative risks were also used. The study proves that both radon exposure and cigarette smoke leads to higher lung cancer risks [10].

III. METHODOLOGY

A database consisting of 309 record sets of patients who visited a website in August 2013 to participate in a survey that tries to predict the possibility of occurrence of lung cancer based on values given for 16 attributes has been used for the research (online lung cancer prediction system).

The attributes taken for the study are Gender, Age, Smoking, Yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, consumption of alcohol, coughing, shortness of breath, swallowing difficulty, chest pain and occurrence of lung cancer. Out of 309, 39 patients were not suffering from lung cancer and 270 were suffering from lung cancer.

Statistical analysis of the given data base was done using pivot tables in Excel. Apriori analysis was done manually with values derived from pivot tables. Weka Data Mining Tool was used to classify the given data set using decision tree classifiers and to generate significant attributes using Subset Evaluators.

IV. RESULTS OBTAINED

Statistical Analysis yielded the following results:

1. Almost 60% of the people with positive diagnosis of lung cancer are aged above 60.
2. 56% of females and 60% of males with lung cancer are smokers.
3. 57% of the control group has yellow fingers out of which 52.8% have lung cancer. Females have yellow fingers more than males.
4. Female members with anxiety and lung cancer are almost twice as those without anxiety but with lung cancer. Out of those with lung cancer in the entire control group (87%) we find that more than half have anxiety.
5. Peer pressure is more among females than males for those patients who have lung cancer.
6. People with lung cancer have more fatigue, especially men.
7. 54% of our control group suffers from allergies and lung cancer.
8. 52.8% people in the entire control group suffer from wheezing and lung cancer.
9. Most of the people, especially men (81%) having lung cancer, consume alcohol.
10. People with lung cancer are struck with coughing up to 54.7%
11. 57% of the control group has shortness of breath and lung cancer. Men have more shortness of breath than women.
12. Women with lung cancer have more swallowing difficulty (45.3% of the population)
13. Chest pain is found mostly in males with lung cancer (75% of men)

Table 1 Apriori analysis using single item sets

Sino	Symptom	F(Symptom, Lung Cancer)	Support F(Symptom, Lung Cancer) / 309	Confidence F(Symptom, Lung Cancer) / F(Symptom)	Rank in Confidence
1	Fatigue	189	61.2%	90.8%	11
2	Shortness of Breath	176	57%	88.9%	13
3	**Coughing**	169	54.7%	94.4%	5
4	**Allergy**	167	54%	97.1%	1
5	**Alcohol**	165	53.4%	95.9%	3

6	Yellow fingers	163	52.8%	92.6%	8
7	**Wheezing**	161	52.8%	94.8%	4
8	Chest pain	160	51.8%	93%	7
9	Smoking	155	50.2%	89%	12
10	**Peer_ pressure**	145	46.9%	93.5%	6
11	Anxiety	142	46%	92.2%	9
12	Chronic Disease	142	46%	91%	10
13	**Swallowing Difficulty**	130	45.3%	96.6%	2

PREDICTION OF LUNG CANCER USING KNOWLEDGE MINED THROUGH APRIORI:

Apriori analysis of the dataset using 2 itemsets and 3 itemsets provided us with the knowledge that Allergy, Swallowing difficulty, Alcohol consumption, Wheezing, Coughing, and Peer Pressure are significant symptoms that contribute to the occurrence of lung cancer in patients. Even though smoking is generally considered to be the cause for cancer, it did not rank well in our sample data set.

Apriori analysis is a data mining technique used to generate decision rules for determining a result. This technique was applied to the above dataset to generate three decision rules. Based on these decision rules three data models were built.

1. If allergy = "yes" or swallowing difficulty = "yes" or coughing = "yes" or wheezing = "yes" or Alcohol consumption = "yes" then lung cancer = "yes"

Using the above prediction model we got a prediction accuracy of **87%**

2. If allergy = "yes" and swallowing difficulty = "yes" and (coughing = "yes" and wheezing = "yes") and Alcohol consumption = "yes" then lung cancer = "yes"

Using the above prediction model we got a prediction accuracy of **90.9%**

3. If allergy = "yes" and swallowing difficulty = "yes" and (coughing = "yes" or wheezing = "yes") and Alcohol consumption = "yes" then lung cancer = "yes"

Using the above prediction model we got a prediction accuracy of **92.6%**

Table 2 Comparison of Prediction accuracy of Classifiers

Classifier	Correctly Classified Instances	Relative Absolute Error	Roc Area	TP Rate		FP Rate	
	In %	In %		Yes	No	Yes	No
Random forest	91.2	54.46	0.943	0.96	0.56	0.44	0.04
Random tree	90.9	39.98	0.774	0.96	0.56	0.44	0.04
J48 pruned tree	90.3	56.81	0.787	0.96	0.54	0.46	0.04
Naïve bayes	87.4	100	0.488	1.0	0.0	1.0	0.0
Decision table	86.4	90.6	0.841	0.94	0.3	0.69	0.05

V. DISCUSSION

Subset evaluation by WEKA confirmed that the same set of attributes (Allergy, Swallowing difficulty, Alcohol consumption, Wheezing, Coughing, and Peer Pressure) obtained through statistical methods were more predominant than the other attributes in patients with lung cancer. CFS Subset evaluator's best first search method and info gain attribute evaluator's ranker method were used for this.

Apriori analysis of the data set led to the generation of three different rules for predicting the possibility of occurrence of lung cancer in a patient. Each rule led to the formation of a different data model for prediction.

We have used the three different models from the knowledge gained and in each case, we get prediction accuracy as 87%, 90.9% and 92.6%. The third model seems to be the best prediction model as it gives us the maximum accuracy of 92.6%.

The classification of the given dataset using different classifiers led to the following results.

Out of all classifiers RANDOM FOREST classifier gave the highest prediction accuracy of 91.2% with a good True Positive rate. RANDOM TREE and J48 PRUNED TREE classifiers were also closely accurate while the others were far below.

Comparing the maximum accuracy obtained through classification using decision tree models and the accuracy obtained through apriori analysis, we find that apriori analysis has produced a prediction model with accuracy greater than the decision tree classifiers.

Thus it is now possible to predict if a person has a high probability of having lung cancer by watching out for specific symptoms like Allergy, Swallowing difficulty,

Alcohol consumption, Wheezing, Coughing, and Peer Pressure.

VI. CONCLUSION

In this research study, we have taken a database consisting of 309 records of patients reporting various ailments suspected to have lung cancer. Out of 309, 39 patients were not suffering from lung cancer and 270 were suffering from lung cancer. An analysis of the dataset using statistical techniques using pivot tables in Excel, Apriori Analysis and Weka Data Mining tool for subset evaluation and classification revealed a wealth of facts. Knowledge regarding significant factors that are biomarkers for the occurrence of lung cancer was revealed. These were then used in a prediction model to determine the probability of occurrence of lung cancer in patients. The training set was taken to evaluate the prediction model and an accuracy of 92.6% was achieved in prediction. Thus data mining techniques were successfully used as proposed for achieving the objectives.

REFERENCES

[1]. M.Venkat Dass, M.A.Rasheed, M.M.Ali, "*Classification of lung cancer subtypes by data mining technique*", Proceedings of the 2014 International Conference on Control, Instrumentation, Energy and Communication, pages 558-562, ISBN: 978-1-4799-2044-0, Jan 31 2014

[2]. Juliet Rani Rajan, C.Chilambu Chelvan, "*A survey on mining techniques for early lung cancer diagnosis*", Proceedings of the 2013 International Conference on Gren Computing, Communication and Conservation of Energy, ISBN: 978-1-4673-6126-2, 12-14 Dec 2013

[3]. Ahmed K. Emran AA., Jesmin T, Mukti RF, Rahman MZ, Ahmed . "*Early detection of lung cancer risk using data mining",* Asian Pacific Journal of Cancer Prevention, Issue 14(1), Pages 595-598, 2013.

[4]. Ada, Rajneet Kaur, "A Study of Detection of Lung Cancer using Data Mining Classification Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3(3), ISSN 2277 128X, Pages 131-134, Mar 2013

[5]. Haofan Yang, Yi-Phing Phoebe Chen, *"Data Mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information*", Expert Systems with Applications, Vol 42(15), Pages 6168-6176, Sep 2015

[6]. Dr.P.Thangaraju, G.Barkavi, "*Lung cancer early diagnosis using some data mining classification techniques: A Survey*", CompuSoft- An International Journal of Advanced Computer Technology, ISSN 2320-0790, Vol 3(6), June 2014

[7]. Kawsar Ahmed, Tasnuba Jesmin, et al., "*An early Detection of Lung Cancer Risk Using Data mining*", Proceedings of the Bangladesh Society For Biochemistry and molecular Biology Conference 2013, 12-13 Jan 2013

[8]. Guoxin Tang, "*Data mining and Analysis of lung cancer data*", Ph.D Thesis, Dept. Of Mathematics, University of Louisville, USA https://doi.org/10.18297/etd/1418

[9]. V.Krishnaiah, Dr G. Narasimha, Dr.N.Subhash Chandra, "*Diagnosis of Lung Cancer Prediction System using Data Mining Classification Techniques*", International Journal of Computer Science and Information Technologies, Vol 4(1), Pages 39-45, ISSN 0975-9646, 2013

[10]. Suresh H. Moolgavkar, E.Georg Luebeck, Daniel Krewski and Jan M.Zielinski, "*Radon, Cigarette Smoke, and Lung Cancer*" *A Re-analysis of the Colorado Plateau Uranium Miners' Data*", Journal of Epidemiology, ISSN 1044 3983, Vol 4(3),pages 204-217, May 1993

[11]. M. Shukla, A. K. Malviya, "*Analysis and Comparison of Classification Algorithms for Student Placement Prediction*", International Journal of Computer Sciences and Engineering, Vol.6, Issue.6, pp.69-81, 2018.

[12]. Ramisetti Uma Maheswari, R Raja Sekhar, "*Pruning and Ranking Based Classifier for Efficient Detection of Android Malware*", International Journal of Computer Sciences and Engineering, Vol.6, Issue.6, pp.201-205, 2018.

Authors Profile

Mrs. Vijayalakshmi completed her U.G and P.G in Computer Science at Seethalakshmi Ramaswami College, Trichy, Tamilnadu India in 1993 and did her M.Phil in Computer Science in Manonmaniam Sundaranar University, Tirunelveli. She is serving as Associate Professor in Department of Computer Science, Shrimati Indira Gandhi College, Trichy, for the past 24 years. She has published 12 research papers in reputed international journals indexed in Scopus, ICI, Google Scholar etc. Her main research area is data mining and knowledge discovery. She is also interested in Network Security. She has 15 years of research experience.

Ms.J Polly Amilya is currently doing her M.Phil in Computer Science under the guidance of Ms. N. Vijayalakshmi. Her area

