RESEARCH ARTICLE

Early Detection of Lung Cancer Risk Using Data Mining

Kawsar Ahmed[1], Abdullah-Al-Emran[2]*, Tasnuba Jesmin[1], Roushney Fatima Mukti[2], Md Zamilur Rahman[1], Farzana Ahmed[3]

Abstract

Background: Lung cancer is the leading cause of cancer death worldwide Therefore, identification of genetic as well as environmental factors is very important in developing novel methods of lung cancer prevention. However, this is a multi-layered problem. Therefore a lung cancer risk prediction system is here proposed which is easy, cost effective and time saving. Materials and Methods: Initially 400 cancer and non-cancer patients' data were collected from different diagnostic centres, pre-processed and clustered using a K-means clustering algorithm for identifying relevant and non-relevant data. Next significant frequent patterns are discovered using AprioriTid and a decision tree algorithm. Results: Finally using the significant pattern prediction tools for a lung cancer prediction system were developed. This lung cancer risk prediction system should prove helpful in detection of a person's predisposition for lung cancer. Conclusions: Most of people of Bangladesh do not even know they have lung cancer and the majority of cases are diagnosed at late stages when cure is impossible. Therefore early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy.

Keywords: Data mining - pre-processing - disease diagnosis - aprioriTid algorithm - DT algorithm - Bangladesh

Asian Pacific J Cancer Prev, **14 (1)**, 595-598

Introduction

Lung cancer is the most common cause of cancer death worldwide. The occurrence of lung cancer has increased rapidly and become the most common cancer in men in most countries. Lung cancer accounts for around 1,095,000 new cancer cases and 951,000 deaths each year in men, and 514,000 cases and 427,000 deaths in women, representing about 12.7% of all new cancer cases each year and 18.2% of cancer deaths (Ferlay et al., 2010; Paul et al., 2011). Uncontrolled cell growth causes diseases that are known as cancer. Lung cancer occurs for out-of-control cell growth and begins in one or both lungs. Lung cancer that spreads to the brain can cause difficulties with vision, weakness on one side of the body. Symptoms of primary lung cancers include cough, coughing up blood, chest pain, and shortness of breath.

Cigarette smoking is the most important cause of lung cancer. Cigarette smoke contains more than 4,000 chemicals, many of which have been identified as causing cancer. A person who smokes more than one pack of cigarettes per day has a 20-25 times greater risk of developing lung cancer than someone who has never smoked. About 90% of lung cancers arise due to tobacco use (Smith et al., 2012). However, other factors, such as environment pollution mainly air; excessive alcohol may also be contributing for Lung Cancer (Schmid et al., 2010).

Among the overall population of Bangladesh, lifetime mortality risks (per 100,000 population) of cancer of the lung was 159.1, 23.1 for males and females respectively. The prevalence is increasing at an alarming rate in a developing country like Bangladesh in recent years (Ferlay et al., 2010). Therefore the early diagnosis of Lung cancer is obvious but the diagnosis is costly in the developing countries. Therefore based on different and most common risk factors of lung cancer a risk prediction system of lung cancer is proposed in this study which will be cost effective and easy to use.

A widely recognized formal definition of data mining can be defined as "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data". Data mining has some fields to analysis of data such as classification, clustering, correlations, association rule etc (Jayalakshmi and Santhakumaran, 2010) and has been used intensively and extensively by many organizations. And In-healthcare, data mining is becoming increasingly popular. Data mining provides the methodology and technology to analysis the useful information of data for decision making.

*[1]Department of Information and Communication Technology, [2]Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Tangail, [3]Department of Mathematics and Natural Science, BRAC University, Dhaka, Bangladesh *For correspondence: emrangeb@gmail.com, emranbge@mbstu.ac.bd*

Data pre-processing is a vital task of data mining. It mainly used for making analysis appropriate and also making data appropriate for clustering by avoiding duplicate records and adding missing data according to past recorded data. The main benefits of data pre-processing reduces memory.

Clustering is a process of separating dataset into subgroups according to the unique feature. Clustering separated the dataset into relevant and non-relevant dataset to Lung Cancer. AprioriTid (Lan et al., 2010) and Decision Tree algorithm (Yael and Elad, 2010) are mainly used to find out frequent patterns of dataset. Those algorithms are very easy and effective to find out frequent patterns. Frequent patterns, the sets of data are frequently occurred into data warehouse. Significant frequent pattern, the set of data are mostly responsible to Lung Cancer. Using this significant pattern we implemented a prediction system for Lung Cancer.

The main goal of this research is to develop a system that can be used by a person for testing his/her Lung Cancer risk level.

Materials and Methods

400 patients' data (200 lung cancer patients and 200 non-cancer patients) is obtained from different diagnostic centre. There are 200 male and 200 female patients whose age between 20-80 years old. From the previous studies 20 risk factors were considered for Lung cancer assessment in Bangladeshi population, which includes-age, gender, hereditary, previous health examination, use of anti-hypersensitive drugs, smoking, food habit, physical activity, obesity, tobacco, genetic Risk, environment, mental trauma, uptake of red meat, balance diet, hypertension, heart disease, excessive alcohol, radiation therapy and chronic lung diseases.

Data pre-processing is a vital term of data mining. Making an appropriate analysis and suitable for clustering of collected data. This is the main goal of data pre-processing. Sometimes data warehouse is consisted with duplicate data and missing any values of data. Data pre-processing deletes the duplicates data and supplies the missing values according to the past recorded data. It also reduces the memory and normalizes the values used to represent information in database.

The process of partitioning and category of collected data into different subgroups where each groups have a unique feature is called clustering. Clustering is another tedious term of data mining. The clustering problem has been addressed in numerous contents besides being proven beneficial in many applications (Muhammad et al.,2011). The goal of clustering is to classify objects or data into a number of categories or classes where each class contains identical feature. The main benefits of clustering are that the data object is assigned to an unknown class that have unique feature and reduces the memory.

The K-means clustering (Amorim and Mirkin, 2012) is a widely recognized clustering tool that is used for robotics, diseases and artificial intelligence application purposes (Pradhan and Kumar, 2011). Here k is a positive integer representing the number of clusters. The pre-processed data is clustered using the K-means clustering algorithm with the value of k equal to 2. This represents there is two clusters where one cluster contains relevant data to Lung Cancer and another contains remaining data that means non relevant data.

This is the most significant and vital topics of data mining. It is considered as the principle data mining problem that intends to find out the frequent items or patterns from the data warehouse. There are different kinds of algorithms, used to mine interesting frequent patterns from databases like association rules, clusters, classifications and correlations etc such as Apriori, AprioriTid, Decision Tree, and FP-Tree.

After clustering, AprioiTid (Lan et al., 2010) and Decision Tree algorithms (Yael and Elad, 2010) is used to mine the frequent patterns. The AprioriTid and Decision Tree algorithms are the efficient algorithms of extracting the frequent patterns from clustered dataset.

Pseudo code for Algorithm AprioriTid and Decision tree

1) L1 = {large 1-itemsets}; 2) For (k = 2; Lk ≠ Φ; k++) do begin; 3) Ck = apriori-gen(Lk≠1); // New candidates; 4) Forall transactions t € D do begin; 5) Ck = Φ; 6) Forall entries t €Ck-1 do begin; 7) // determine candidate itemsets in Ck contained in the transaction with identi_er t.TID Ct = {c € Ck | (c -c[k]) € t:set-of-itemsets ^ (c-c[k-1]) € t.set-of-itemsets}; 8) forall candidates c € Ct do; 9) c.count++; 10) If (Ct ≠ Φ) then Ck += < t:TID,Ct >; 11) End; 12) Lk = {c € Ck | c.count ≥minsup}; 13) End; 14) Answer = Uk Lk;

Input II: is a set of candidate attributes, and **S** is a set of labelled instances

Output: A decision Tree **T**; 1) **If** (**S** is pure or empty) or (**II** is empty) **Return T**. 2) Compute Ps(Ci) on **S** for each class Ci. 3) **For** each attribute X in **II**, compute IIG(S,X) based on equation 1 and 5. 4) Use the attribute **X**max with the highest IIG for the root. 5) Partition S into disjoint subsets Sx using **X**max. 6) **For** all values x of **X**max •**T**x=NT(**II**-**X**max,**S**x), •Add **T**x as a child of **X**max. 7) **Return T**.

After discovering the frequent patterns using AprioriTid and Decision Tree algorithm, the weightage significant patterns are mined by using the Equation (1)

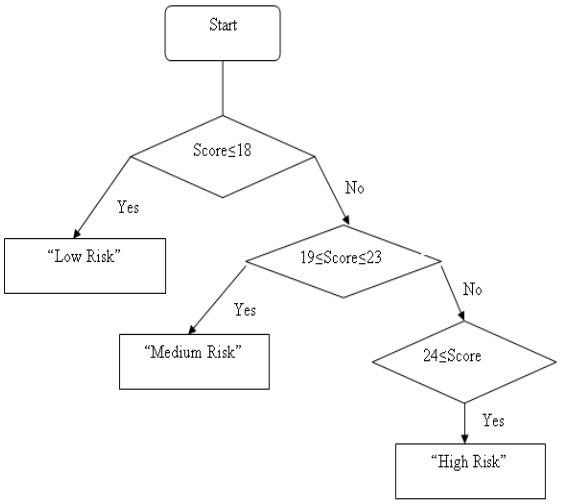

Figure 1. Flow Diagram of Decision Tree Algorithm

Table 1. Significant Pattern and Their Corresponding Weightage Value using AprioriTid Algorithm and Decision Tree Algorithm

Significant Patterns	AprioriTid Algorithm Weight age	Decision Tree Algorithm Weight age
Age-Smoking-Sex-Obesity-Tobacco-Alcohol-Environment-Mental trauma	200.55	200.55
Age-Smoking-Sex-Tobacco-Environment-Genetic Risk-Chronic Lung Disease	180.05	180.05
Age-Balance Diet- Smoking-Sex-Tobacco-Environment –Radiation Therapy	175.50	175.50
Smoking- Sex- Genetic Risk-Tobacco- Mental trauma–Radiation Therapy	160.55	160.55
Smoking-Obesity- Environment--Chronic Lung Disease-Balance Diet-Mental trauma	155.05	155.05

Table 2. Significant Pattern and Their Corresponding Weight Age and Score

Parameters	Weight age			Score
Age	≤40			1
	<40-≤60			2
	>60			3
Sex	Male			3
	Female			1
Air Pollution	No			1
	Yes			2
Excessive alcohol use	No			1
	Yes			2
Radiation therapy to chest area				
	No			1
	Yes			2
Occupational hazard	No			1
	Yes			2
Genetic risk	No			1
	Yes			2
Chronic lung diseases	No			1
	Yes			2
balanced diet	No			2
	Yes			1
Obesity	No			1
	Yes			2
Tobacco	No			1
	Yes			3
Smoking	Yes			3
	No	Passive Smoker	Yes	2
			No	1

(Gothwal et al., 2011), $Sw(i)=\sum(Wi*Fi)$ (1).

Where W_i is the weightage of each attribute and F^i represents number of frequency for each rule. And significant Frequent Pattern is selected by using the following Equation (2) $SFP=Sw(n) \geq \phi$ for all values of n (2). Where SFP denotes significant frequent pattern and ϕ denotes significant weightage.

Results

The experimental results are separated into two sections. One is significant frequent patterns discover and another is represents prediction tools to Lung Cancer.

Using data from data warehouse, the significant patterns are extracted for Lung cancer prediction. The collected data are pre-processed by deleting duplicate records and adding missing values. Then pre-processed data is clustered using K-means cluster algorithm with k equal to 2. And finally significant frequent patterns are mined using AprioriTid shown in Table 1 and Decision Tree algorithm shown in Table 2.

Figure 2. Lung Cancer Prediction with Low Risk Level

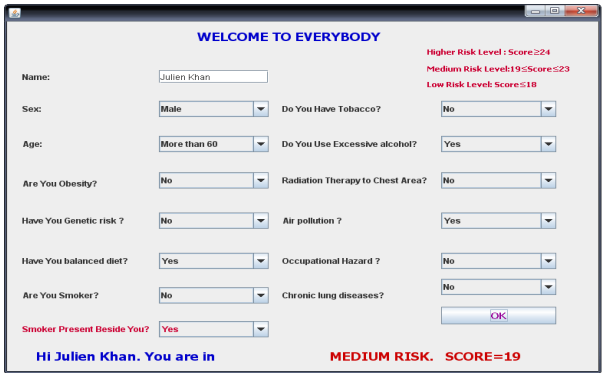

Figure 3. Lung Cancer Prediction with Medium Risk Level

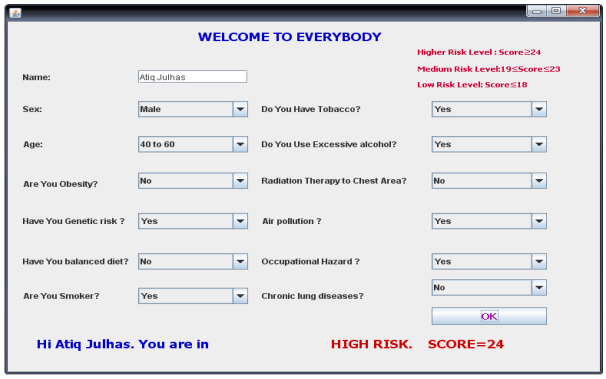

Figure 4. Lung Cancer Prediction with High Risk Level

Finally using the significant pattern the prediction tools to Lung Cancer are implemented. Table 3 represents the frequent pattern parameters and their corresponding score and Figure 1 represents the risk level of Lung Cancer which is implemented using Table 3.

Discussion

Large numbers of people in the Bangladesh and the world have Lung cancer. Most of them do not even

know they have it. There is no remedy for cancer after completely affected. Death is inevitable. So the ability to predict Lung cancer plays an important role in the diagnosis process. In this paper we have proposed an effective Lung cancer prediction system based on data mining. We have provided an efficient approach for the extraction of significant pattern from data warehouse for efficient prediction of Lung cancer. The proposed method is implemented using java. The proposed method can efficiently and successfully predict the risk of Lung cancer.

Acknowledgements

The authors are grateful to the participants who contributed to this research.

References

Amorim R, Mirkin B (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern Recognition*, **45**,1061-75.

Brennan P, Hainaut P, Boffetta P (2011). Genetics of lung-cancer susceptibility. *Lancet Oncol*, **12**, 399-408.

Ferlay J, Shin HR, Bray F, et al (2010). GLOBOCAN 2008: cancer incidence and mortality worldwide: *IARC*, **10**, 220-7.

Gothwal H, Kedawat S, Kumar R (2011). Cardiac arrhythmias detection in an ECG beat signal using fast fourier transform and artificial neural network. *J Bio Sci Engineering*, **4**, 289-96.

Jayalakshmi T, Santhakumaran A (2010). A novel classification method for classification of diabetes mellitus using artificial neural networks. International Conference on Data Storage and Data Engineering. 159-63

Lan C, Liu Y, Tang Z (2010). Improvement of aprioritid algorithm for mining frequent items[J]. *Computer Applications And Software*, **27**, 234-6.

Manaswini P, Ranjit KS (2011). Predict the onset of diabetes disease using artificial neural network (ANN). *Int J Computer Sci & Emerging Technologies*, **2**, 303-11.

Muhammad ASapon, Khadijah Ismail, Suehazlyn Zainudin (2011). Prediction of diabetes by using artificial neural network. 2011 International Conference on Circuits, System and Simulation, **7**, 299-303.

Schmid K, Kuwert T, Drexler H (2010). Radon in indoor spaces: an underestimated risk factor for lung cancer in environmental medicine. *Dtsch Arztebl Int*, **107**, 181-6.

Smith L, Brinton LA, Spitz MR, et al (2012) Body mass index and risk of lung cancer among never, former, and current smokers. *J Natl Cancer Inst*, **104**, 778-89.

Yael Ben-Haim , Elad Tom-Tov (2010) A streaming parallel decision tree algorithm. *J Machine Learning Res*, **11**, 849-72.