Volume 7, Issue 5, May 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

General Ideas for Using Data Mining Techniques in Lung Cancer

Dr. A. R. Pon Periasamy
Associate Professor of Computer Science
Nehru Memorial College
Puthanampatti, Trichy (DT)
Tamilnadu, India

K. Arutchelvan
Assistant Professor / Programmer
Department of Pharmacy
Annamalai University, Chidamparam
Tamilnadu, India

DOI: 10.23956/ijarcsse/V7I4/0165

Abstract— ***Data mining intends to endow with a systematic survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research. Tobacco smoke Smoking is by far the leading risk factor for lung cancer. In the early lung cancer was much less common than some other types of cancer. But these changed once manufactured cigarettes became readily available and more people began smoking. At least 80% of lung cancer deaths are thought to result from smoking. The risk for lung cancer among smokers is many times higher than among non-smokers. The longer you smoke and the more packs a day you smoke, the greater your risk. Cigar smoking and pipe smoking are almost as likely to cause lung cancer as cigarette smoking. Smoking low-tar or "light" cigarettes increases lung cancer risk as much as regular cigarettes. There is concern that menthol cigarettes may increase the risk even more since the menthol allows smokers to inhale more deeply. It strives to improve software quality and testing efficiency by constructing predictive models from code attributes to enable a timely identification of fault-prone modules. The main objective of paper is to help developers identify defects based on existing software metrics using data mining techniques and thereby improve the software quality. In this paper, we will discuss data mining techniques that are association mining, classification and clustering for lung cancer prediction.***

Key Words: ***Lung cancer, data mining, clustering, classification and association rule mining.***

I. INTRODUCTION

Cancer is a class of diseases characterized by out-of-control cell growth. There are over 100 different types of cancer, and each is classified by the type of cell that is initially affected.

1.1 Causes of Cancer

Cancer is ultimately the result of cells that uncontrollably grow and do not die. Normal cells in the body follow an orderly path of growth, division, and death. Programmed cell death is called apoptosis, and when this process breaks down, cancer begins to form. Unlike regular cells, cancer cells do not experience programmatic death and instead continue to grow and divide. This leads to a mass of abnormal cells that grows out of control.

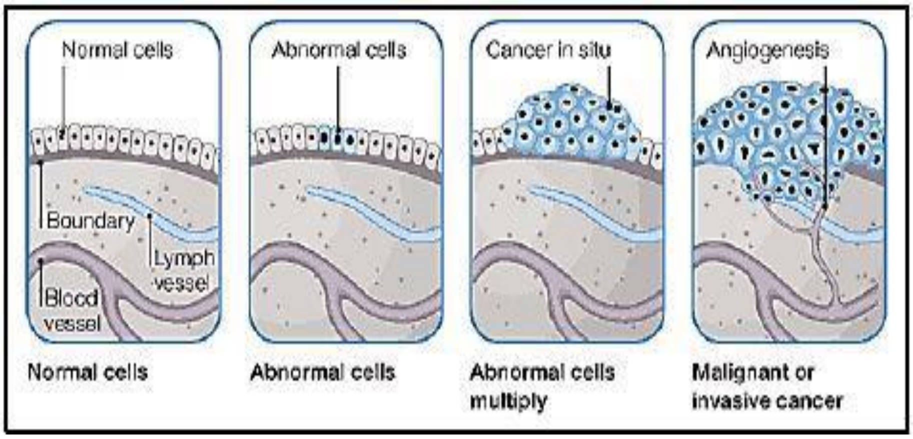

Figure 1.2 Cancer Cell

More dangerous, or malignant, tumors form when two things occur:

1. a cancerous cell manages to move throughout the body using the blood or lymph systems, destroying healthy tissue in a process called invasion
2. That cell manages to divide and grow, making new blood vessels to feed itself in a process called angiogenesis.

When a tumor successfully spreads to other parts of the body and grows, invading and destroying other healthy tissues, it is said to have metastasized. This process itself is called metastasis, and the result is a serious condition that is very difficult to treat.

1.2 Paths of cancer spread

Scientists reported in *Nature Communications* (October 2012 issue) that they have discovered an important clue as to why cancer cells spread. It has something to do with their adhesion (stickiness) properties. Certain molecular interactions between cells and the scaffolding that holds them in place (extracellular matrix) cause them to become unstuck at the original tumor site, they become dislodged, move on and then reattach themselves at a new site.

The researchers say this discovery is important because cancer mortality is mainly due to metastatic tumors, those that grow from cells that have traveled from their original site to another part of the body. Only 10% of cancer deaths are caused by the primary tumors.

The scientists, from the Massachusetts Institute of Technology, say that finding a way to stop cancer cells from sticking to new sites could interfere with metastatic disease, and halt the growth of secondary tumors.

In 2015, cancer claimed the lives of about 7.6 million people in the world. Physicians and researchers who specialize in the study, diagnosis, treatment, and prevention of cancer are called oncologists.

1.3 Symptoms of cancer

Cancer symptoms are quite varied and depend on where the cancer is located, where it has spread, and how big the tumor is. Some cancers can be felt or seen through the skin - a lump on the breast or testicle can be an indicator of cancer in those locations. Skin cancer (melanoma) is often noted by a change in a wart or mole on the skin. Some oral cancers present white patches inside the mouth or white spots on the tongue.

As cancer cells use the body's energy and interfere with normal hormone function, it is possible to present symptoms such as fever, fatigue, excessive sweating, anemia, and unexplained weight loss. However, these symptoms are common in several other maladies as well. For example, coughing and hoarseness can point to lung or throat cancer as well as several other conditions.

When cancer spreads, or metastasizes, additional symptoms can present themselves in the newly affected area. Swollen or enlarged lymph nodes are common and likely to be present early. If cancer spreads to the brain, patients may experience vertigo, headaches, or seizures. Spreading to the lungs may cause coughing and shortness of breath. In addition, the liver may become enlarged and cause jaundice and bones can become painful, brittle, and break easily. Symptoms of metastasis ultimately depend on the location to which the cancer has spread.

1.4 Common Types of Cancer

Skin cancer is the most commonly diagnosed cancer among men and women. Over one million cases are diagnosed each year, with more young people having skin cancer than ever before. The most common types of cancer on frequency of diagnosis are:

- lung cancer
- breast cancer
- bladder cancer
- colon cancer
- endometrial cancer
- kidney cancer (renal cell)
- leukemia
- melanoma
- non-Hodgkin lymphoma
- pancreatic cancer
- prostate cancer
- thyroid cancer

1.5 Cancer diagnosing and staging

Early detection of cancer can greatly improve the odds of successful treatment and survival. Physicians use information from symptoms and several other procedures to diagnose cancer. Imaging techniques such as X-rays, CT scans, MRI scans, PET scans, and ultrasound scans are used regularly in order to detect where a tumor is located and what organs may be affected by it. Doctors may also conduct an endoscopy, which is a procedure that uses a thin tube with a camera and light at one end, to look for abnormalities inside the body.

Figure 1.3 Cell Diagnosing

Extracting cancer cells and looking at them under a microscope is the only absolute way to diagnose cancer. This procedure is called a biopsy. Other types of molecular diagnostic tests are frequently employed as well. Physicians will analyze your body's sugars, fats, proteins, and DNA at the molecular level. For example, cancerous prostate cells release a higher level of a chemical called PSA (prostate-specific antigen) into the bloodstream that can be detected by a blood test. Molecular diagnostics, biopsies, and imaging techniques are all used together to diagnose cancer.

After a diagnosis is made, doctors find out how far the cancer has spread and determine the stage of the cancer. The stage determines which choices will be available for treatment and informs prognoses. The most common cancer staging method is called the TNM system. T (1-4) indicates the size and direct extent of the primary tumor, N (0-3) indicates the degree to which the cancer has spread to nearby lymph nodes, and M (0-1) indicates whether the cancer has metastasized to other organs in the body. A small tumor that has not spread to lymph nodes or distant organs may be staged as (T1, N0, M0), for example. TNM descriptions then lead to a simpler categorization of stages, from 0 to 4, where lower numbers indicate that the cancer has spread less. While most Stage 1 tumors are curable, most Stage 4 tumors are inoperable or untreatable.

1.6 Cancer treatments

Cancer treatment depends on the type of cancer, the stage of the cancer (how much it has spread), age, health status, and additional personal characteristics. There is no single treatment for cancer, and patients often receive a combination of therapies and palliative care. Treatments usually fall into one of the following categories: surgery, radiation, chemotherapy, immunotherapy, hormone therapy, or gene therapy.

1.6.1 Surgery

Surgery is the oldest known treatment for cancer. If a cancer has not metastasized, it is possible to completely cure a patient by surgically removing the cancer from the body. This is often seen in the removal of the prostate or a breast or testicle. After the disease has spread, however, it is nearly impossible to remove all of the cancer cells. Surgery may also be instrumental in helping to control symptoms such as bowel obstruction or spinal cord compression.

1.6.2 Radiation

Figure 1.4 Radiotherapy

Radiation treatment, also known as radiotherapy, destroys cancer by focusing high-energy rays on the cancer cells. This causes damage to the molecules that make up the cancer cells and leads them to commit suicide. Radiotherapy utilizes high-energy gamma-rays that are emitted from metals such as radium or high-energy x-rays that are created in a special machine. Early radiation treatments caused severe side-effects because the energy beams would damage normal, healthy tissue, but technologies have improved so that beams can be more accurately targeted. Radiotherapy is used as a standalone treatment to shrink a tumor or destroy cancer cells (including those associated with leukemia and lymphoma), and it is also used in combination with other cancer treatments.

1.6.3 Chemotherapy

Chemotherapy utilizes chemicals that interfere with the cell division process - damaging proteins or DNA - so that cancer cells will commit suicide. These treatments target any rapidly dividing cells (not necessarily just cancer cells), but normal cells usually can recover from any chemical-induced damage while cancer cells cannot. Chemotherapy is generally used to treat cancer that has spread or metastasized because the medicines travel throughout the entire body. It is a necessary treatment for some forms of leukemia and lymphoma. Chemotherapy treatment occurs in cycles so the body has time to heal between doses. However, there are still common side effects such as hair loss, nausea, fatigue, and vomiting. Combination therapies often include multiple types of chemotherapy or chemotherapy combined with other treatment options.

1.6.4 Immunotherapy

Immunotherapy aims to get the body's immune system to fight the tumor. Local immunotherapy injects a treatment into an affected area, for example, to cause inflammation that causes a tumor to shrink. Systemic immunotherapy treats the whole body by administering an agent such as the protein interferon alpha that can shrink tumors. Immunotherapy can also be considered non-specific if it improves cancer-fighting abilities by stimulating the entire immune system, and it can be considered targeted if the treatment specifically tells the immune system to destroy cancer cells. These therapies are relatively young, but researchers have had success with treatments that introduce antibodies to the body that inhibit the growth of breast cancer cells. Bone marrow transplantation (hematopoetic stem cell transplantation) can also be considered immunotherapy because the donor's immune cells will often attack the tumor or cancer cells that are present in the host.

II. DATA MINING

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large database in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation (Tan P., Steinbach M. and Kumar V. 2006). Data mining is a method of extracting what is useable within a database and separating it out from what is unusable. Such methods are necessary because, as human being, we lack the capacity to sort and organize such large volumes of data.

The quick growth in data and databases has created a pressing need for new tools and techniques that can quickly and efficiently process raw data into useable information. In the last years the development of information technology has motivated a parallel growing of facilities to store and manage database. The largest amount of stored data is more important for the demand of extracting the implicit information they contained to aid the decision-making in business, health care services, research…etc. Thus, to obtain useful knowledge from the data stored in large repositories (i.e. the "knowledge discovery"), which is recognized as a basic necessity in many area. Since nineties of the last century, the research area named "data mining" has become central topic in databases and Artificial Intelligence. The task of discovering association rules was introduced by [Agrawal, Imieliński and Swami 1993]. In its original form, the task was defined for a special kind of data, often called basket data, where a tuple consists of a set of binary attributes called items. Each tuple corresponds to a customer transaction, where a given item has a value of true or false, depending on whether or not the corresponding customer bought the item in that transaction. This kind of data is usually collected through bar-code technology; the typical example is a supermarket scanner.

There are various algorithms have been proposed to discover the frequent item sets (Sarawagi, Thomas and Agrawal 1998), (Agrawal, Imieliński and Swami1993). The Apriori algorithm is one of the most popular algorithms in the mining of association rules in a centralized database, which will explained broadly later.

III. BACKGROUND OF WORK

Lung cancer is a cancer that starts in the lungs. To understand lung cancer, it helps to know about the normal structure and function of the lungs. The lungs your lungs are two sponge-like organs found in your chest. Your right lung is divided into three sections, called lobes. Your left lung has two lobes. The left lung is smaller because the heart takes up more room on that side of the body.

When you breathe in, air enters through your mouth or nose and goes into your lungs through the trachea (windpipe). The trachea divides into tubes called the bronchi (singular, bronchus), which enter the lungs and divide into smaller bronchi. These divide to form smaller branches called bronchioles. At the end of the bronchioles are tiny air sacs known as alveoli. Many tiny blood vessels run through the alveoli. They absorb oxygen from the inhaled air into your bloodstream and pass carbon dioxide from the body into the alveoli. This is expelled from the body when you exhale. Taking in oxygen and getting rid of carbon dioxide are your lungs' main functions.

A thin lining layer called the pleura surrounds the lungs. The pleura protects your lungs and helps them slide back and forth against the chest wall as they expand and contract during breathing. Below the lungs, a thin, dome-shaped muscle called the diaphragm separates the chest from the abdomen. When you breathe, the diaphragm moves up and down, forcing air in and out of the lungs.

3.1 Start and spread of lung cancer

Lung cancers can start in the cells lining the bronchi and parts of the lung such as the bronchioles or alveoli. Lung cancers are thought to start as areas of pre-cancerous changes in the lung. The first changes in the genes (DNA) inside the lung cells may cause the cells to grow faster. These cells may look a bit abnormal if seen under a microscope, but at this point they do not form a mass or tumor. They cannot be seen on an x-ray and they do not cause symptoms. Over time, the abnormal cells may acquire other gene changes, which cause them to progress to true cancer. As a cancer develops, the cancer cells may make chemicals that cause new blood vessels to form nearby. These blood vessels nourish the cancer cells, which can continue to grow and form a tumor large enough to be seen on imaging tests such as x-rays. At some point, cells from the cancer may break away from the original tumor and spread (metastasize) to other parts of the body. Lung cancer is often a life-threatening disease because it tends to spread in this way even before it can be detected on an imaging test such as a chest x-ray. Types of lung cancer. There are two main types of lung cancer:

- Small cell lung cancer (SCLC)
- Non-small cell lung cancer (NSCLC)

3.2 Small cell lung cancer

About 10% to 15% of all lung cancers are small cell lung cancer (SCLC), named for the size of the cancer cells when seen under a microscope. Other names for SCLC are oat cell cancer, oat cell carcinoma, and small cell undifferentiated carcinoma. It is very rare for someone who has never smoked to have small cell lung cancer. SCLC often starts in the bronchi near the center of the chest, and it tends to spread widely through the body early in the course of the disease. This cancer is discussed in the document Lung Cancer (Small Cell).

3.3 Non-small cell lung cancer

About 85% to 90% of lung cancers are non-small cell lung cancer (NSCLC). There are three main subtypes of NSCLC. The cells in these subtypes differ in size, shape, and chemical make-up. But they are grouped together because the approach to treatment and prognosis (outlook) are often very similar.

3.4 Squamous cell (epidermoid) carcinoma

About 25% to 30% of all lung cancers are squamous cell carcinomas. These cancers start in early versions of squamous cells, which are flat cells that line the inside of the airways in the lungs. They are often linked to a history of

3.5 Adenocarcinoma

About 40% of lung cancers are adenocarcinomas. These cancers start in early versions of the cells that would normally secrete substances such as mucus. This type of lung cancer occurs mainly in current or former smokers, but it is also the most common type of lung cancer in non-smokers. It is more common in women than in men, and it is more likely to occur in younger people than other types of lung cancer. Adenocarcinoma is usually found in the outer parts of the lung. It tends to grow slower than other types of lung cancer, and is more likely to be found before it has spread outside of the lung. People with a type of adenocarcinoma called adenocarcinoma in situ (previously calledbronchioloalveolar carcinoma) tend to have a better outlook (prognosis) than those withother types of lung cancer.

3.6 Large cell (undifferentiated) carcinoma

This type of cancer accounts for about 10% to 15% of lung cancers. It can appear in any part of the lung. It tends to grow and spread quickly, which can make it harder to treat. A subtype of large cell carcinoma, known as large cell neuroendocrine carcinoma, is a fast-growing cancer that is very similar to small cell lung cancer.

3.7 Lung carcinoid tumors

Carcinoid tumors of the lung account for fewer than 5% of lung tumors. Most are slow-growing tumors that are called typical carcinoid tumors. They are generally cured by surgery. Some typical carcinoid tumors can spread, but they usually have a better prognosis than small cell or non-small cell lung cancer. Less common are atypical carcinoid tumors. The outlook for these tumors is somewhere in between typical carcinoids and small cell lung cancer. For more information about typical and atypical carcinoid tumors, see the document Lung Carcinoid Tumor.

3.8 Other lung tumors

Other types of lung cancer such as adenoid cystic carcinomas, lymphomas, and sarcomas, as well as benign lung tumors such as hamartomas are rare. These have different risk factors from the more common lung cancers.

3.9 Cancers that spread to the lungs

Cancers that start in other organs (such as the breast, pancreas, kidney, or skin) can sometimes spread (metastasize) to the lungs, but these are not lung cancers. For example, cancer that starts in the breast and spreads to the lungs is still breast cancer, not lung cancer. Treatment for cancer that has spread to the lungs is based on which type of cancer it is.

3.10 The risk factors for lung cancer

A risk factor is anything that affects a person’s chance of getting a disease such as cancer. Different cancers have different risk factors. Some risk factors, like smoking, can be changed. Others, like a person’s age or family history, can’t be changed. But risk factors don’t tell us everything. Having a risk factor, or even several risk factors, does not mean that you will get the disease. And some people who get the disease may not have had any known risk factors. Even if a person with lung cancer has a risk factor, it is often very hard to know how much that risk factor may have contributed to the cancer. Several risk factors can make you more likely to develop lung cancer.

IV. CONCLUSION

To achieve medical data of higher quality all the necessary steps must be taken in order to build the better medical information systems which provides accurate information regarding to patients medical history rather than the information regarding to their billing invoices. Because high quality healthcare data is useful for providing better medical services only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry. All the necessary steps are takes in order to minimize the semantic gap in data sharing between distributed healthcare databases environment so that meaningful patterns can be obtained. These patterns can be very useful in order to improve the treatment effectiveness services, to better detection of fraud and abuse, improved customer relationship management across the world.

The privacy regarding to patient’s confidential information is very important. Such type of privacy may be lost during sharing of data in distributed healthcare environment. Necessary steps must be taken in order to provide proper security so that their confidential information must not be accessed by any unauthorized organizations. But in situations like epidemic, planning better healthcare services for a very large population etc. some confidential data may be provided to the researchers and government organizations or any authorized organizations. In order to achieve better accuracy in the prediction of diseases, improving survivability rate regarding serious death related problems etc. various data mining techniques must be used in combination. It is believed that the data mining can significantly help in the Lung Cancer research and ultimately improve the quality of health care of Lung Cancer patients. It can also be implemented using several classification techniques. Future models can be used in the design of clinical decision support system for mining Lung Cancer.

REFERENCES

[1] Osareh. A and Shadgar. B, "Microarray data analysis for cancer classification", 5th International Symposium on Health Informatics and Bioinformatics (HIBIT), 2010.

[2] Kai-Lin Tang, Wei-Jia Yao, Tong-Hua Li, Yi-Xue Li And Zhi-Wei Cao, ―Cancer Classification From The Gene Expression Profiles By Discriminant Kernel-Pls‖,Journal Of Bioinformatics And Computational Biology,Vol.8,Suppl.1(2010) 147-160.

[3] A. Bharathi And Dr.A.M.Natarajan,"Cancer Classification Of Bioinformatics Data Using ANOVA ", International Journal Of Computer Theory And Engineering, Vol. 2, No. 3, 1793-8201,2010.

[4] R. Mallika, And V. Saravanan, "An SVM Based Classification Method For Cancer Data Using Minimum Microarray Gene Expressions", World Academy Of Science, Engineering And Technology 62 2010.

[5] N. Revathy And R. Amalraj, "Accurate Cancer Classification Using Expressions Of Very Few Genes", International Journal Of Computer Applications, Vol.14, No.4, 2010.

[6] Ritu Chauhan “Data clustering method for Discovering clusters in spatial cancer databases” International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

[7] Rajashree Dash “A hybridized K-means clustering approach for high dimensional dataset” International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66.

[8] G. Rajkumar “ Intelligent Pattern Mining and Data Clustering for Pattern Cluster Analysis using Cancer Data” International journal of Engineering Science and Technology Vol. 2(12), 2010, ISSN: 7459-7469

[9] Santanu Ghorai, Anirban Mukherjee, Sanghamitra Sengupta, And Pranab K. Dutta, " Cancer Classification From Gene Expression Data By NPPC Ensemble", IEEE/Acm Transactions On Computational Biology And Bioinformatics, Vol. 8, No. 3, May/June 2011.

[10] K.Kalaivani “Childhood Cancer-a Hospital based study using Decision Tree Techniques” Journal of Computer Science 7(12): 1819-1823, 2011 ISSN: 1549-3636

[11] Shyi-Ching Liang, Yen-Chun Lee and Pei-Chiang Lee , “The Application of Ant Colony Optimization to the Classification Rule Problem”, 2011 IEEE International Conference on Granular Computing.

[12] Arezoo Modiri and Kamran Kiasaleh,” Permittivity Estimation for Breast Cancer Detection Using Particle Swarm Optimization Algorithm”, 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011

[13] Yao Liu and Yuk Ying Chung, “Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning”, IEEE 2011.

[14] Hnin Wint Khaing,” Data Mining based Fragmentation and Prediction of Medical Data”, IEEE 2011.

[15] M. H. Mehta,” Hybrid Genetic Algorithm with PSO Effect for Combinatorial Optimisation Problems”, International Journal of Advanced Computer Research (IJACR), Volume-2 Number-4 Issue-6 December- 2012.

[16] Priyanka Dhasal, Shiv Shakti Shrivastava, Hitesh Gupta, Parmalik Kumar, “An Optimized Feature Selection for Image Classification Based on SVMACO”, International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-3 Issue-5 September-2012.

[17] Amog Rajenderan,” Data Preparation for Web Mining A survey”, International Journal of Advanced Computer Research (IJACR),Volume-2 Number-4 Issue-6 December-2012.

[18] Pragati Shrivastava,Hitesh Gupta, “A Review of Density-Based clustering in Spatial Data”, International Journal of Advanced Computer Research (IJACR) ,Volume-2 Number-3 Issue-5 September-2012.

[19] Boris Milovic “Prediction and Decision Making in Health Care using Data Mining” International Journal of Public Health Science Vol. 1, No. 2, December 2012, pp. 69-78 ISSN: 2252-8806