ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

Cancer Spread Pattern – an Analysis using Classification and Prediction Techniques

P.Ramachandran[1], Dr.N.Girija[2], Dr.T.Bhuvaneswari[3]

Ph.D Research Scholar, Computer Science & Application Department, SCSVMV University, Enathur, Kanchipuram, Tamil Nadu, India[1]

Asst.Professor, Information Technology Department, Higher College of Technology, Ministry of Manpower, Muscat[2]

Asst.Professor, Computer Science & Application Department, L. N. Govt.of Arts & Science College, Ponneri, Chennai, Tamil Nadu, India[3]

Abstract: Cancer is one of the dreadful diseases in the world claiming majority of lives. The aim of this research is to find out the cancer spread pattern in rural and urban areas and districts in South India and classify them according to age, sex, blood group, habits, education, types of cancer etc using various data mining techniques such as classification and prediction with WEKA tool. We use Linear Regression for the task of prediction with which we can predict the areas and blood groups that have the most vulnerability towards cancer in future. We can also predict the most common types of cancer that occur among men and women. This research may provide valuable information to enhance the cancer control program.

Keywords: Cancer, Linear Regression, Data Mining, Classification, Prediction

I. INTRODUCTION

A. Problem: Cancer has become the leading cause of death worldwide. The most effective way to reduce cancer deaths is to detect it earlier. Though cancer research is generally clinical and biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques have to be applied. Prevention of tobacco-related and cervical cancers and earlier detection of treatable cancers would reduce cancer deaths in India. The absolute number of cancer deaths in India is projected to increase due to population growth and increasing life expectancy.

In this paper we preprocess a data set which contain 17 attributes and approximately 3000 cases of cancer recorded in each of the following years, 1998, 1999, 2000 and 2001 and classify them with data mining techniques and find out which sex is most affected by cancer and what are the top three types of cancers in male and female cases and which blood group is more prone to cancer. We also implement accuracy in this classification by finding out the ratio of cancer in male versus female cases and also which areas record highest cancer rate and whether habits, diets, education, marital status, living area etc., play important roles in cancer pattern. To this classification we append the results of Linear Regression using the method of least squares and predict the recorded outcome for the year 2002. The *Prediction Appender* takes a Classifier and a dataset and appends the classifier's predictions to the dataset. As a next step we preprocess a dataset for the year 2002 and cross examine the classification results with our predictions and analyze whether our predictions are accurate and whether machine learning truly helps in cancer prediction.

B. Cancer among Children

The rate at which cancer cases increases among kids is alarming. Among 40000-50000 cases which are detected in India every year, 6000 cases are detected in Tamil Nadu alone. Inadequate cancer treatment facilities, lack of pediatric oncologists, high costs of treatment and repeated painful procedures are some of the reasons for the inability to treat more cases. Recent studies show that there is a direct link between rapid urbanization and increase in the number of cancer cases among children. "There is definitely an increased recognition of childhood cancer, a willingness of people to come forward for diagnosis, this could well have translated into more cases, but there is an overall trend of an increase in cancer with urbanization," says Dr AnanthaKrishnan pediatric oncologist. Treatment of cancer especially in kids needs a lot more in terms of financial and emotional support. "There is no research on childhood

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

cancer in India and hardly any trained staff in hospitals. There are few support groups for families," says Dr Julius Scott, Pediatric Hemato oncologist, SRMC. In this paper we give more importance to cancers that occur among children and find out what are the most common childhood cancers.

II. RELATED WORK

Rajaraman Swaminathan et al [1] Cancer incidence was significantly lower, cancer patterns were markedly different and population-based cancer survival was lower in rural areas than urban areas thus providing valuable leads in estimating realistic cancer burden. Rajaraman Swaminathan et al [3] with more and more women in rural India becoming educated, one could foresee breast cancer becoming more frequent even in rural areas of India in future. Shweta Kharya et al [13] discuss some of effective techniques that can be used for breast cancer classification. Among the various data mining classifiers and soft computing approaches, Decision tree is found to be best predictor on benchmark dataset (UCI machine learning dataset) and also on SEER dataset. T.Sakthimurugan et al [11] keyword searching algorithm used to retrieve relevant healthcare information for the corresponding user symptoms and the KNN classifier are used to classify the semantic relations between disease and treatment. Neha Sharma et al [12] suggested an ED&P Framework which is an information system is presented that will deliver the necessary information to clinical, administrative, and policy researchers and analysts in an effective and efficient manner.

A.Sudha et al [14] gives an idea about major life-threatening diseases and their diagnosis using data mining with minimum number of attributes and creates awareness about diseases which leads to death. K.Balachandran et al [2] Early detection of the cancer disease is crucial in diagnosing and treating the patient. Hence it is very essential that common man who has some symptoms and risk factors are better to undergo medical examination by a specialist at the earliest.

The Emphasis of this work is to find the target group of people who needs further screening for Lung cancer disease, so that the prevalence and mortality rate could be brought down. Jaree Thongkam et al [7] the OOS approach has been proposed and applied to the tasks of building accurate breast cancer survivability prediction models. It also states that difficult to choose an appropriate method for developing prediction models. S. Aruna et al [22] to find out the best classifier with respect to accuracy, precision, sensitivity and specificity in detecting breast cancer. Jin Oh Kang et al [5] predicts the hospital charge incurred by the cancer patients. The ANN models showed better prediction accuracy than CART models. However, the CART models, which serve different information from ANN model, can be used to allocate limited medical resource effectively and efficiently.

III. FACTORS

The following are the most common risk factors for developing cancer. They are, Growing older, Tobacco, Sunlight, Ionizing radiation, Certain chemicals and other substances, Some viruses and bacteria, Certain hormones, Family history of cancer, Alcohol, Poor diet, lack of physical activity, or being overweight. We found out some cancers are strongly related to carcinogenic agents such as tobacco smoking or chewing are oral (including lip, oral, and pharynx), lung (including trachea and larynx), mouth, esophagus, bladder, kidney, throat, stomach, pancreas, or cervix. They also are more likely to develop acute myeloid leukemia (cancer that starts in blood cells). People who use smokeless tobacco (snuff or chewing tobacco) are at increased risk of getting oral cancer. Cervical cancer (human papilomavirus), stomach cancer, liver (hepatitis B and C) and associated cancers are mostly related to Infections and viruses. Human T-cell leukemia/lymphoma virus (HTLV-1): Infection with HTLV-1 increases a person's risk of lymphoma and leukemia. Having more than two drinks of alcohol per day for many years may increase the chance of developing cancers of the mouth, throat, esophagus, larynx, liver, and breast. The risk increases with the amount of alcohol that a person drinks. For most of these cancers, the risk is higher for a drinker who uses tobacco. We observed from our experiment that most cancers occur in people above the age of 40.

A. Education

From our experiment we found out cancer rate is lesser in people who are educated above secondary grade and more in people who are uneducated. This may be due to less awareness, unhygienic habits, unhealthy lifestyle, having more number of sexual partners, unsafe sex etc. From our data we conclude 78% of people who have cancer are uneducated or school drop-outs. Education is inversely associated with cancer in men and women. Cancers of mouth, lung esophagus are found high among illiterate people. We also observed from our data that breast cancer is prevalent among educated women than uneducated women. This may be due to high stress level in jobs, lack of time to concentrate on themselves, lack of self physical examination of the breast etc. We also found out that women who gave birth in younger age are less prone to breast cancer than women who gave birth in their later 20's and 30's. Anyhow we found out that marital status has no relationship with cancer as very few people remain unmarried for greater period of time.

B. Diets

Studies have linked consumption of red or processed meat to an increased risk of breast cancer, colon cancer, and pancreatic cancer, a phenomenon which could be due to the presence of carcinogens in foods cooked at high temperature. We found out that people who consume non-

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

vegetarian food are more likely to be affected by cancer than people who consume vegetarian food. This phenomenon is may be due to vegetarian food which includes fruits and vegetables contain high fiber and substances which can prevent cancer. Eating fast food also results in stomach cancer.

C. Living area

From our research we found out that people living in urban areas are more likely to be affected by cancer than those living in rural areas due to environmental factors. Exposure to asbestos, a group of minerals found in housing and industrial building materials can cause a variety of medical problems. Studies have shown that people who are exposed to high amount of benzene are prone to cancer. Benzene is a chemical found in gasoline, smoking, and pollution. Ultraviolet (UV) radiation comes from the sun, sunlamps, and tanning booths causes early aging of the skin and skin damage that can lead to skin cancer.

D. Blood group

In our dataset most of the cancer patients belong to the blood group O (+ve). This may be due to the fact that most of the people throughout the world have O (+ve) as their blood group. It is followed by B (+ve), A (+ve), AB (+ve), B (–ve), A (–ve), and O (–ve) in decreasing order. For example in the year 1998 out of 3699 cancer records 1307 records belong to O (+ve), 1234 records belong to B (+ve), 617 records belong to A (+ve), 399 records belong to AB (+ve), 71 records belong to B (–ve), 36 records belong to A (–ve), and 35 records belong to O (-ve). This pattern is observed in all the five years records including the year 2002.

IV. METHODOLOGY

A. Preprocessing

Preprocessing consumes the biggest portion of work. We took approximately 3000 cancer cases in each year for our study i.e., approximately 15000 cases altogether. Originally we had 30000 records generously provided by Adyar Cancer Research Institute for our research work which contained many missing values and noise. We preprocessed these records to get approximately 3000 records per year. We filtered these records year wise and subsequently from each year we classified the cases age wise. Microsoft Excel and SPSS tool were used for this work. We split the cases age wise in the range of 1-10, 11-20… 90-100.

B. Classification

Classification is a process of finding a model that describes and distinguishes data classes or concepts, for the purpose arranging the records into different class labels and also to predict the class of objects whose class label is unknown. Using WEKA tool's preprocessing and classification methods we classified the cancer records in each age group into different types of cancer. The ICD code that corresponds to different types of cancer is pre-defined. The different types of cancer are our class labels according to which we classify our data. This classification lets us to predict some needed results for our study and we append these results to our classification. This is known as Prediction Appender.

C. Prediction

The most widely used approach for prediction is Regression. In our research we used Linear Regression for our prediction task. As stated above we classified records into different age groups and different types of cancer. By Prediction Appender we appended the results that we obtained from classification to the dataset. Thus by doing this for each of the year 1998, 1999, 2000 and 2001 taken for consideration we can predict which types of cancer incidence will be high in the year 2002, also the highest rate of cancer will be observed in which age group and what are the top three types of cancer that can affect children and elders.

D. Linear Regression

For linear Regression we used the formula,

$$Y = a + bX$$

Where, **a** and **b** are constants and **Y** is response variable and a single predictor variable **X**. We found these constants using the Method of Least Squares as linear regression follows a straight line path. To find out the constants we use two more equations,

$$\sum y = na + b\sum x \quad -1$$

$$\sum xy = a\sum x + b\sum x^2 \quad -2$$

E. Observations

From the all four years of data we classified the record that is below 10 years of age as child cancer. Among all the different types of cancer that is in the age range of 1-10 the top three cancers are Leukemia (myeloid, lymphoma), followed by lymphoma (non Hodgkin, Hodgkin) and Eye cancer. Thus children and babies are mostly affected by leukemia which is evident from the following table.

Table: 1

ICD	1998	1999	2000	2001
Leukemia	15	17	13	18
Lymphoma	12	10	10	14
Eye	5	7	6	12
Bone	4	5	5	4

From this table we calculate and predicted the results for the year 2002 as below

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

Table: 2

ICD	Leukemia	Lymphoma	Eye	Bone
2002	17	13	12	4

Using linear regression we compiled the above table for the year 2002. Based on the above results we expected the number of cancer cases in children in the year 2002 to be 73 out of 3250 cases. Age group between 11-20 is the crucial period in which children attain puberty as well as they fall into bad company acquiring bad habits such as smoking and drinking which have greater impact during old age and results in different types of cancer.

According to our results this is the period during which bone marrow cancer instances are in a steep rise. Later teen age i.e. 17-25 is the age during which people tend to be sexually active and from the age of 21 starts the first dominant instances of cervical cancer but then leukemia tends to occupy the first position followed by breast and cervical cancers. Age specified cancer risk is higher in the age group of 40-60 years. This may be due to growth factor as well as the average life expectancy of an Indian is 60 years and people tend to die of other factors before cancer is diagnosed.

Table: 3

Years	1998	1999	2000	2001
Observed frequency	1577	1600	1750	1651

This table gives us the number of cancer instances between the age of 40 and 60 years. In 2002 the expected frequency is 1737. The observed frequency is 1721. The top three types of cancer in women are Cervix followed by breast, stomach and other gynecology related cancer such as ovarian cancer, vaginal cancer etc. In men the top three cancers are lungs, stomach and esophagus. Above 60 years there is a decline of cancer rates in men and women this may be due to most of the people die of recurring cancer before they reach the age of 70 or due to some other factors like heart attack etc. cancers of cervix, breast esophagus and lungs again occupy the top positions in the age range of 70-100 years. Thus we can see that prediction yields approximate results in our cancer dataset which is more or less similar to our observed results and we can use linear regression for our other prediction works.

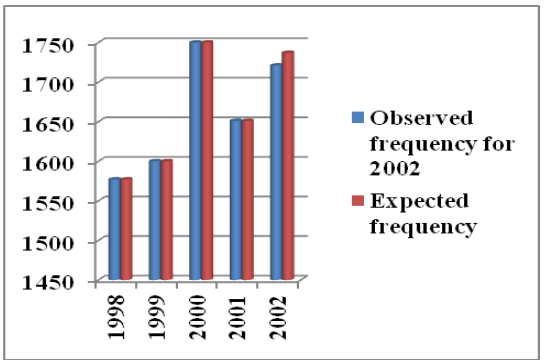

Fig.1 Example of an Observed Frequency for 2002

V. EXPERIMENTAL RESULTS

A. Classification Results

From our experiment we found out that age specified cancer incidence is highest in the age group of 40 to 60 years and women are more prone to cancer than men in almost all age groups. We also conclude external factors such as living in urban area, educational status and habits such as smoking, consuming alcohol chewing tobacco products play a major role in the occurrence of cancer. People who belong to the blood group O (+ve) also have the higher risk of getting cancer followed by B (+ve) and A (+ve) respectively. Marital status does not play a vital role in causing cancer however child birth plays a significant role. Non-vegetarians have more risk of getting cancer than vegetarians. Family history of cancer should be considered as an important factor in diagnosing cancer.

B. Prediction Results

We analyzed the four years 1998, 1999, 2000, 2001 data and using Linear Regression we predicted the outcome for the test dataset in the year 2002. From our expected results and observed results we conclude that Regression provides approximate results and not accurate results. Yet regression could be used to predict the outcome of following years in a similar manner to know the approximate number of cancer cases in every age group since it is highly difficult to obtain a accurate result for forth coming years and rate of cancer is fluctuated due to various causes.

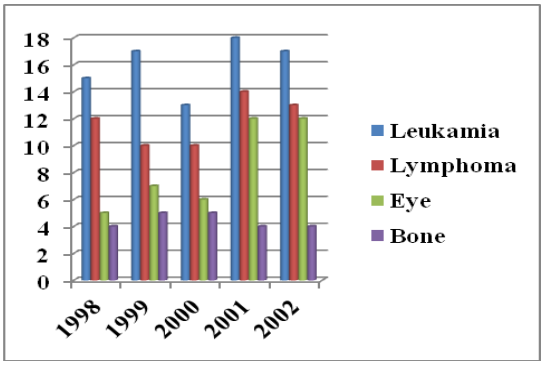

Fig.2 Example of a Prediction for 2002

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

VI. CONCLUSION

This research work provides a valuable knowledge on cancer spread pattern and its factors. This previously unknown knowledge is mined from the dataset provided to us from the Data Warehouse of Adyar Cancer Research Institute. We also conclude that Linear Regression can be used for Prediction tasks as it yields more approximate results.

In future we would like to analyze a latest dataset comprising of last five years and compare the results of this paper with the newly obtained results and to know how the cancer spread pattern is present before and after 10 years.

This research provides valuable knowledge to plan and enhance the cancer control program in India efficiently. It also stresses the point that Child cancer should be controlled and the needed steps should be taken immediately.

ACKNOWLEDGMENT

My sincere thanks to Dr.P.T.Vijayashree, Principal and Lecturers of Kumararani Meena Muthiah College (Co-Ed) Adyar, Ch-20 for motivating and helping me. My heartfelt thanks to Adyar Cancer Institute (WIA), for providing facility for data collection Work. Dr.R.Swaminathan and Dr.V.Shanta for providing valuable suggestions in my research work.

REFERENCES

[1] Rajaraman Swaminathan "Cancer pattern and survival in a rural district in South India" Cancer Epidemiology 33 (2009) 325–331.

[2] K.Balachandran, Dr. R.Anitha "Supervised Learning Processing Techniques for Pre-Diagnosis of Lung Cancer Disease" ©2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4

[3] Rajaraman Swaminathan "Education and cancer incidence in a rural population in south India" Cancer Epidemiology 33 (2009) 89–93.

[4] E. Barati "A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction" Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI): March Edition, 2011.

[5] Jin Oh Kang "Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques" 2009;15(1):13-23.

[6] Thair Nu Phyu "Survey of Classification Techniques in Data Mining" Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[7] Jaree Thongkam "Toward breast cancer survivability prediction models through improving training space" Expert Systems with Applications 36 (2009) 12200–12209.

[8] Zakaria Suliman Zubi "Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer" ISBN: 978-960-474-273-8.

[9] Dr. DSVGK Kaladhar "Data mining, inference and prediction of Cancer datasets using learning algorithms" International Journal of Science and Advanced Technology (ISSN 2221-8386) Volume 1 No 3 May 2011 http://www.ijsat.com

[10] Shelly Gupta "Performance Analysis of Various Data Mining Classification Techniques on healthcare Data" International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011.

[11] T.Sakthimurugan "An Effective Retrieval of Medical Records using Data Mining Techniques" International Journal of Pharmaceutical Science and Health Care, Issue 2, Volume 2 (April 2012) ISSN: 2249-5738.

[12] Neha Sharma "Framework for Early Detection and Prevention of oral Cancer Using Data Mining" International Journal of Advances in Engineering & Technology, Sept 2012. ©IJAET ISSN: 2231-1963, Vol. 4, Issue 2, pp. 302-310.

[13] Shweta Kharya "Using Data Mining Techniques For Diagnosis and Prognosis of Cancer Disease" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012, DOI: 10.5121/ijcseit.2012.2206.

[14] A.Sudha "Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability" International Journal of Computer Applications (0975 – 8887) Volume 41– No.17, March 2012.

[15] Shelly Gupta "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis" Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2011.

[16] S.Vijiyarani & S.Sudha "Disease Prediction in Data Mining Technique – A Survey" www.ijcait.com International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)

[17] V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45, ISSN : 0975-9646.

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

[19] Dr.Varun Kumar "Binary Classifiers for Health Care Databases: A Comparative Study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer" IJCST Vol. 1, Issue 2, December 2010, ISSN : 2229-4333(Print) | ISSN : 0976-8491 (Online)

[20] Remco. R. Bouckaert "Weka-Experiences with a Java Open-Source Project" Journal of Machine Learning Research 11 (2010) 2533-2541.

[21] Dr. Medhat Mohamed Ahmed Abdelaal " Using data mining for assessing diagnosis of breast cancer" Proceedings of International MultiConference on Computer Science and Information Technology pp.11-17, ISBN 978-83-60810-27-9 ISSN 1896-7094.

[22] S. Aruna " Knowledge Based Analysis of Various Statistical Tools in Detecting Breast Cancer" CCSEA 2011, CS&IT 02, pp. 37-45, 2011.

[23] Amir Fallahi "An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network" International Journal of Advanced Science and Technology Vol. 34, September 2011.

[24] P.Ramachandran, Dr. N. Girija, Dr. T. Bhuvaneswari "Healthcare Service Sector: Classifying and Finding Cancer spread pattern in Southern India using Data Mining Techniques" International Journal of Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 4 No. 05 May 2012.

[25] N. Revathy "Accurate Cancer Classification using Expressions of Very few Genes" International Journal of Computer Applications (0975 – 8887) Volume 14– No.4, January 2011.