Healthcare Service Sector: Classifying and Finding Cancer spread pattern in Southern India Using Data Mining Techniques

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

[1]Ph.D Research Scholar, Computer Science & Application Department, SCSVMV University, Kanchipuram, Tamil Nadu, India
rkmvc.rc@gmail.com

[2]Asst.Professor, Information Technology Department, Higher College of Technology, Ministry of Manpower, Muscat
nbgir2004@gmail.com

[3]Asst.Professor, Computer Application Department, Govt. of Arts & Science College for Women, Bargur, Krishnagiri, Tamil Nadu, India
t_bhuvaneswari@yahoo.com,

ABSTRACT

Data mining has evolved from an experimental analysis to a predicting methodology due to highly precise algorithms and high performance data mining tools. Knowledge discovery in database has been used to predict survivability and diagnosis of diseases in the field of medicine. This can prove helpful for prevention of epidemics. In this paper, J48 algorithm classification is used for classifying data based on cancers, patient's gender, age marital status and education using WEKA tool and the pattern of spread of cancer is discussed.

Keywords: *Classification, Cancer Pattern, Spread Of Cancer, Age Based Spread, Weka.*

INTRODUCTION

Classification, a data mining task is an effective method to classify the data in the process of Knowledge Data Discovery. A Classification method, Decision tree algorithms are widely used in medical field to classify the medical data for diagnosis.The advent of newer and easy to use technologies, compact storage devices and development of software indirectly boosted the development of data mining. Data, which is the base of data mining, started following at higher rate and ignited the need to find newer pattern and store them instead of huge data set. These patterns produced further lead to pattern mining. Classification has been identified as important problem in data mining [1]. Medicine is an age old field which contains higher complexities and data rather than any other field. So immediately after the data mining started becoming important, the field of medicine served the resource needed. Data mining on medical data can help in simple classification to highly accurate predictions. The advantage over using classification on medical data would be to get over all idea of the data based on various attributes, so that the complexity can be reduced and detection of anomalies becomes easier.

Knowledge discovery is highly important on any medical data because they can help in detecting a spread of epidemic, categorizing the pandemics using attributes like the locality of the victims, age based spread, gender based spreading etc..Cancer is one such disease that has wider range of spread in India. Statistically, India is found to have higher rate of increase in cancer patients.

With the rapid advancement in information technology, many different data mining techniques and approaches have been applied to complementary medicines[4].Cancer data has higher complexities due to various types of cancer and various method of diagnosis. But there is also need to classify them based on age, sex, marital status of the patient, education qualification so that newer pattern can be found.

ORGANIZATION OF THIS PAPER

This paper contains general introduction about cancer and their spread in southern districts of India and then discussed about classification based on types of Cancer, District, Diagnosis, Method of Diagnosis, age, gender, education and marital status of the patient and later how these attributes affect the spread of cancer is discussed.

RELATED WORK

Zakaria Suliman Zubi, Rema Asheibani et al [5] Lung cancer is a disease of uncontrolled cell growth in tissues of the lung, Lung cancer is one of the most common and deadly diseases in the world. Detection of lung cancer in its early stage is the key of its cure. In general, a measure for early stage lung cancer diagnosis mainly includes those utilizing X-ray chest films, CT, MRI, etc. Rajaraman swaminathan, Ramanujam selvakumaran and pulikattil okkuru esmy et al [2] 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 and instituting cancer control programs in India. R Swaminathan et al [7] 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. Jin Oh Kang, Suk-Hoon Chung and Yong-Moo Suh et al [9] to predict both the total amount of hospital charges and the amount paid by the insurance of cancer patients and compared their efficacies. V Shanta, G Selvaluxmy, R Swaminathan, P Shanthi et al [13] our data indicates satisfactory treatment outcome even in advanced disease and with the present state of knowledge, the recommended standard treatment for LACC is careful pre-treatment evaluation followed by CCRT which includes brachytherapy.

A. Cancer Spread In Southern India

In southern region, especially around Tamil Nadu and Andhra Pradesh, the cancer details of registered cancer patients were found out in from the data acquired from Adyar cancer institute, Chennai. The data provided contains details such as age, diagnostic status, and method of diagnosis, sex, district and educational qualification.

The data accuracy was perfect because now a day's all government hospitals up to district level had trained technicians for coding disease using ICD-10 and sub-district hospitals had facilities for storage and retrieval of medical records[2]. The data thus obtained were cleansed. Then the data selection was made so that the analysis can be better and later feed into WEKA data mining tool. The J48 algorithm was used for classification.

B. Classification Methodology

C4.5 is a well known decision tree induction learning technique which has been used by Abdelghani Bellaachia and Erhan Gauven[3].

WEKA tool provides this C4.5 decision tree using J48 algorithm. So the data that were pre processed were provided to the tool in CSV format. The classification was done based on following Attributes:

Age: the age of the registered patient suffering from any type of cancer.

Sex: the gender of the patient

Marital status: unmarried, married, widowed, and divorced.

Educational qualification: illiterate, high school, graduate, scholar.

These factor can been used to find out some rare patterns in the data and if the pattern found is good enough the spread of cancer can be traced from difference point of view than just the type and method of diagnosis.

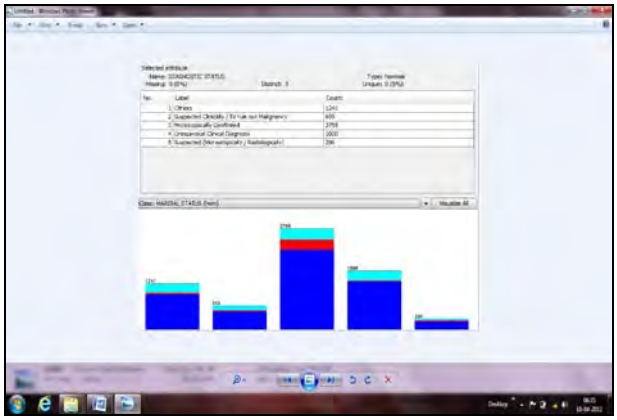

FIGURE 1.1: STATISTICAL BAR GRAPH BASED ON DIAGNOSIS STATUS

C. Classification Using Gender

Gender classification was done to find if cancer has any sex differentiated pattern in spread. Usually there are few regions of cancer spread that are found to be gender differential. Breast cancer and vaginal cancer are to be affected only for females and so to better the classification the gender difference is seemed to be crucial.

So the data got from Adyar cancer institute was used to classify based on gender and how other attributes affect based on gender.

It was found that unmarried males outnumbered females in cancer and widowed and divorced female had much spread of cancer than that of males.

The analysis has little to speak about the cancer spread but the number were denoted to be 416 unmarried male were affected by cancer to that of 137 unmarried females. Also 957 windowed female were diagnosed of cancer to that of 112 males and 57 divorced females were found to have cancer to that of 6 males.

The ratio of unmarried cancer patient of male to that of female was 3.03 and ratio of widowed female to that of male was 8.54 and divorced females versus males was 9.5.This provides an little deeper meaning of how the cancer has spread and since cancer is found to get higher due to stress or depression these factors can be scored out as the unmarried male below age 26 or so are stressed more and widowed female or divorced female are under higher depression. Yet, this cannot be the complete conclusion because there are other factors like age and educational qualifications that can be used to get little bit in-depth in the cancer pattern found in the southern region of India.

The gender based classification based on age was by average the age as 50 and considered for cancer patients below 50 and above age of 50. This also provided little more detail to the classification pattern such that females of age below 50 or equal to it were found to have cancer higher than male, while male outnumbered females after the age of 50.

The reason can be due to the breast cancer because it seems to affect the women from the age of 35 to 50. So the gender based classification proves to much important and crucial for such pattern mining. Now considering this with the previous conclusion it is logical to find that women below the age of 50 has higher cancer spread due to marital status stress or depression. While men seems to have picked up the cancer either below 30 years of age or above 50 years of age due to the same reason.

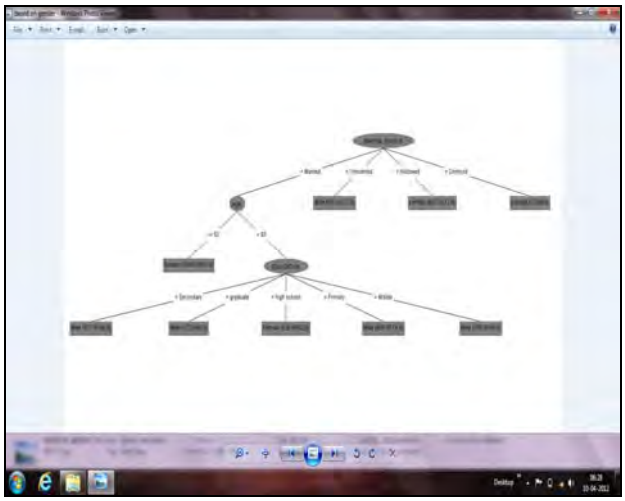

FIGURE 1.2: Classification of Cancer Pattern Based on Gender

D. Classification Using Age

Classification of cancer data based on gender provided little more insight into cancer pattern while there is little more to support the reasons and so there is a need to classify the cancer pattern based on age. The age based classification was a factor to detect how many unmarried female below 26 and married male below 26 has cancer. Seeing so there were almost equal number of married female above age 19 to below 26 affected by cancer to that of unmarried male above the age of 19 to below 26 years of age. This pattern sound to be little insightful to find out that unmarried male and married female till the age of 26 has equal ratio of cancer. This can evident that stress caused to do early marriage for girls can be the reason for spread and the stress cause to workload can be reason for male below 26 can be the cause to cancer spread.

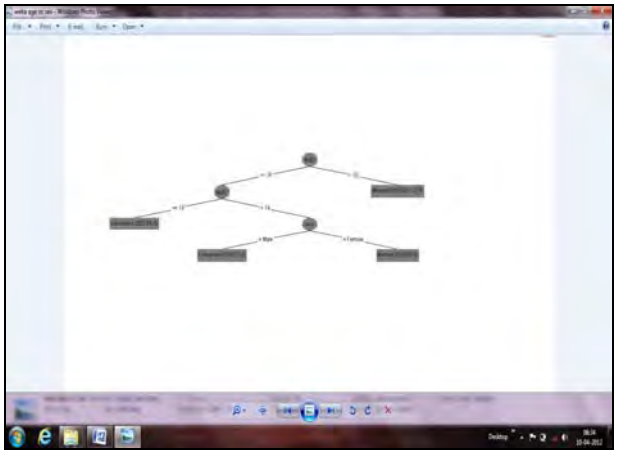

FIGURE 1.3: Classification of Cancer Pattern Based on Age and Gender

Yet the education pattern below and above the average can help in shedding little more light on the topic. It was found that below the age of 50 male with primary, middle school, graduate or higher secondary had cancer more than women. Also female patients registered with high school had cancer more than that of male.

E. Classification Based On Educational Qualification

Educational qualification of patients or victims of cancer was classified based on high school, middle school, graduate, higher secondary or primary. These puts people who are considered illiterate under the primary category and rest in the other category.

The educational qualification classification was done for gender, age and diagnostic status so that the pattern of cancer spread can be analyzed by the educational qualification of the cancer patients.

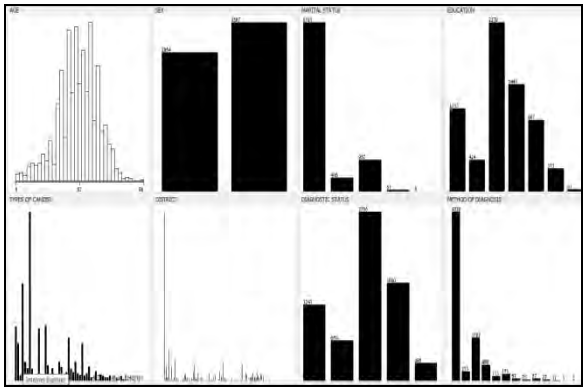

FIGURE 1.4: Statistical Chart of Various Attributes

The above chart denotes statistical chart about the patients with cancer based on various attributes.

It is found that male who is married and had educational qualification of secondary affected due to cancer was 427 and females were married and higher secondary affected due to cancer was just about 133. These numbers had lot to do with the economical and at the same to find the cultural pattern of the cancer victims affected. Deeply analyzing the data it is found most women affected due to cancer where either above 35 to 50 or below 26 which means that early marriage might have been the cause or breast cancer might have cause these outputs.

The figures also show that the separated women who had cancer and diagnosed the disease was much more that of the male which can also add to the fact that women having stress as the reason for higher cancer spread that men having depression had. Logically speaking the spread of cancer is not gender based but the spread of cancer when classified based on the gender get the whole big picture idea that crucial pattern are those of the details about the educational qualification, marital status and gender based classification is helpful for finding and much to say estimating or predicting the spread of cancer in the southern region of Tamil nadu. Since India is diverse land and there are various cultural changes in different parts of the India. This conclusion can't be justified or used for finding or estimating the cancer spread in whole of India but rather could act as the first step to include the simple attributes about the patients to mine the pattern for cancer spread and late for determining the survivability.

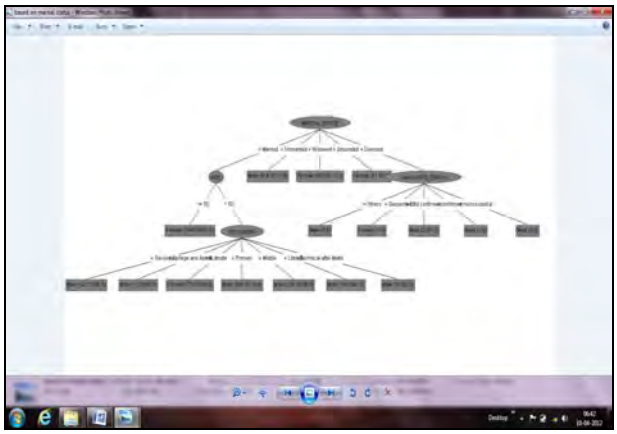

FIGURE 1.5: Education Based Classification of Cancer Pattern

FUTURE RESEARCH

The cancer spread and pattern that was classified has shown that the attributes such as age, gender, education and marital status too are important aspects of cancer pattern and not only that they also act as a crucial factor while predicting the cancer pattern.

Future work on this can be suggested that not only cancer pattern in southern part of India but cancer pattern around the world can be deeply review and researched to find out any anomalies or cause of the cancer or at least determine the flow of it in the way.

CONCLUSION

This paper is just the first step to find pattern using classified data. There is further ways to find cancer pattern to avoid it spread which medical is a miraculous outcome. Cancer is the disease that has been for years and yet incurable due to lot of medical needs and lack of technical advancement.

But in this age of higher technology and greater potential data mining can be helping had only if it can help not only finding newer pattern but using predictive methodologies to predict the future work.

ACKNOWLEDGEMENTS

My sincere thanks to Adyar Cancer Institute (WIA), for providing facility data collection Work. R.Swaminathan and V.Shanta for providing valuable suggestions in my research work.

REFERENCES

[1] Varun Kumar, Nisha Rathee “Knowledge discovery from database using an integration of clustering and classification” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No.3, March 2011, PP.No:29-33.

[2] Rajaraman swaminathan, Ramanujam selvakumaran and pulikattil okkuru esmy, “Cancer pattern and survival in a rural district in south india”, Elsevier ltd, 2009.

[3] Shelly gupta and Dharminder kumar and ananad Sharma, “Data mining classification techniques applied for breast cancer diagnosis and prognosis”,IJCSE, vol.2 april-may 2011.

[4] k.Balachandran and Dr. R.Anitha, “supervisory expert system approach for pre-diagnosis of lung cancer”, IJAEA, January 2010.

[5] Thair Nu Phyu, “survey of classification techniques in data mining”, IMECS 2009.

[6] Zakaria Suliman Zubi, Rema Asheibani Saad “Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer” Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, ISBN: 978-960-474-273-8, PP.No: 32-37.

[7] R Swaminathan “Education and cancer incidence in a rural population in South India”, Cancer Epidemiology 33 (2009) 89–93 Cancer Institute (WIA), Chennai 600036, India.

[8] Alaa M. Elsayad “Diagnosis of Breast Tumor using Boosted Decision Trees” ICGST-AIML Journal, Volume 10, Issue 1, October 2010, PP.No:01-11.

[9] Jin Oh Kang, Suk-Hoon Chung and Yong-Moo Suh “Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques” Journal of Korean Society of Medical Informatics 15-1, 13-23, 2009.

[10] Anil Rajput and Ramesh Prasad Aharwal “Approaches of Classification to Policy of Analysis of Medical Data” International Journals of Computer Science and Network Security, Vol.9 No.11, November 2009 PP.No:209-217.

[11] R.Swaminathan, V.Shanta “Changing Cancer Incidence Pattern, Trend and Future Burden in Chennai, Tamil Nadu, India”

[12] R.Swaminathan, Esmy PO “Dindigul Ambilikkai Cancer Registry: Providing Leads to Rural Cancer Registration and Cervical Cancer Control in Tamil Nadu, India”

[13] V Shanta, G Selvaluxmy, R Swaminathan, P Shanthi "Evolution in the Management of Locally Advanced Cervical Cancer: The Experience of Cancer Institute (WIA), Chennai, India" Management of Locally Advanced Cervical Cancer in the Cancer Institute (WIA), Chennai, India, Asian Pacific Journal of Cancer Prevention, Vol 11, 2010, PP.No: 1091-1097.
[14] S.Arun, Dr.S.P.Rajagopalan and L.V.Nandakishore "Knowledge Based analysis of Various Statistical Tools in Detecting Breast Cancer" Computer Science & Information Technology (CS&IT) 02, PP.37-45, 2011.
[15] Dharminder Kumar & Deepk Bhardwaj "Rise of Data Mining: Current and Future Application Areas" IJCSI Vol.8, Issue 5, No.1, September 2011.
[16] T. Deepa, Dr.M.Punithavalli "Evaluating the Performance of Filtering Techniques for Feature Selection in High Dimensional Imbalanced Dataset" 2010 IEEE International Conference on Computational Intelligence and Computing Research, ISBN: 97881 8371 3627.
[17] Abdullah.H Wahbeh, Qasem A.Al-Radaideh, Mohammed N. Al-Kabi and Emad M.Al-Shawakfa " A Comparison Study between Data Mining Tools over some Classification Methods" International Journals of Advanced Computer Science and Applications(IJACSA). PP: 18-26.
[18] Debahuti Mishra and Barnali Sahu "Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach" International Journal of Scientific & Engineering Research, Volume2, Issue4, April-2011 ISSN 2229-5518 PP.No:01-07.
[19] Dharminder Kumar and Deepak Bhardwaj "Rise of Data Mining: Current and Future Application Areas" International Journal of Computer Science Issues, Vol.8, Issue 5, No 1, September 2011 ISSN (Online):1694-0814, PP.No:256-260.
[20] Abdallah.Z and Gaber Mohamed "KB-CB-N Classification: Towards Unsupervised Approach for Supervised Learning" Proceedings of the IEEE Symposium on Computational Intelligences and Data Mining (CIDM 2011), April 11-15, 2011.
[21] Remco R.Bouckaert "Weka – Experiences with a Java Open Source Project" Journal of Machine Learning Research 11(2010), 2533-2541.
[22] Paulo J.G.Lisboa "Data Mining in Cancer Research" IEEE Computational Intelligence Magazine, February2010, 1556-603.
[23] D.Lavanya & Dr.K.Usha Rani "Analaysis of Feature Selection with Classification: Breast Cancer Datasets" Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166, Vol.2 No.5 Oct-Nov 2011.

AUTHORS PROFILE

Mr.P.Ramachandran M.C.A,M.Sc,M.Phil, Assistant Professor, Department of Computer Science & Computer Applications, Kumararani Meena Muthiah College, Adyar, Chennai-20 then Guest Lecturer in Ramakrishna Mission Vivekananda College, Mylapore, Chennai-14. Currently doing his Ph.D in SCSVMV University then I have published four International Journals and Presented & Participated several Papers in International Conferences.

Dr.N.Girija M.C.A, Ph.D., Currently working as Asst.Professor in Information Technology Department, Higher College of Technology, Ministry of Manpower, Muscat. She is completed her research work on Knowledge Data Mining and Awarded Ph.D degree by SCSVMV University in Year 2007. She is worked as Professor and Head of the Department of MCA, R.M.D.Engineering College, Chennai. She has published Several Papers in both International & National Journals. She is authored three books in Computer Science & Engineering.

Dr.T.Bhuvaneswari M.C.A, Ph.D Assistant Professor in Computer Application Department, Govt. of Arts & Science College for Women, Bargur, Krishnagiri, Tamil Nadu, India. She is completed his research work on Knowledge Data Mining and Awarded Ph.D degree by SCSVMV University in Year 2007. She is worked as Professor and Head of the Department of MCA, Dr.M.G.R University, Chennai. She has published two Papers in an International Journal and Several Papers in National & International Conference. She acted as Panel Member & Chairperson for Several National Conferences. Now, she is Research Supervisor of Bharathiar University & Dr.M.G.R Educational & Research Institute, Chennai.