ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Intelligent Data Mining Techniques in Cancer Disease Pattern Detection

Aditya Mansana [1], Aparna Dubey[2], Dr.Vishnu Kumar Mishra[3]

[1]*Dr.C.V.RAMAN UNIVERSITY, BILASPUR, INDIA,*
MPHIL SCHOLAR

[2]*Dr.C.V.RAMAN UNIVERSITY, BILASPUR, INDIA,*
PHD SCHOLAR

[3]*ASSOCIATE PROFESSOR*
DEPTT.OF ENGINEERING (CSE)
RSR RUNGTA COLLEGE OF ENGINEERING & TECHNOLOGY, BHILAI (C.G)

Abstract-Technology plays a very crucial role for healthcare. Information technologies provides computerized patients records, computerized decision and support tools, computer based hospital information, telemedicine and many more facilities to health care. In general patents records are very huge incomplete, inconsistent, noisy and contains many attributes some of attributes of them may be irrelevant. It is very difficult to analyze huge amount of medical data and take decision and support by manually and Information technologies provide several tools by which it becomes easy to analyze medical data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and interpretation of data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD)[1], refers to as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data".

Keywords-Pattern evaluation Knowledge representation, Fuzzificztion, cancer disease, healthcare

I. INTRODUCTION

The iterative process [2] consists of the following steps:

1. Data cleaning: Also known as data cleansing, it is a phase in which noisy data and irrelevant data are removed from the collection.
2. Data integration: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. Data mining: It is the crucial step in which clever techniques are applied to extract data Patterns potentially useful.
5. Pattern evaluation: In this step, strictly interesting patterns representing knowledge are identified based on given measures.
6. Knowledge representation: It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results. Information technology provides a platform by which it becomes easy to store huge amount of data process the data. Information technology improves the accessibility of data. When health information technology is designed and applied properly it reduces the cost of care, save lives and improves the quality of services [3]. Lots of research works are going on from last few decades in healthcare. Several approaches are proposed to identify hidden patterns in various

diseases. Applying information technology in healthcare becomes very fruitful for our society.

II. PROBLEM IDENTIFICATION

Goal of this research to provide better diagnosis and early diagnosis of cancer patients and preventive measure before situation become critical for cancer patients. Details of patient are provided to propose tool and this tool determine whether the patient belongs from low, medium or high cancer class. Research work will study the impact of various attributes like gender, location place; age and lymph node determine the interdependency between the attributes.

III. LITERATURE SURVEY

Data set is taken from NANAVATI HOSPITAL, Mumbai. First the preprocessing is performed to improve the quality of data and make the data complete, consistent and noiseless. Then perform vertical clustering of data set and group the data set in to two clusters one cluster contains all relevant attributes and other cluster contains irrelevant attributes. Further we process only those records having relevant attributes [4]. Then perform horizontal clustering and make three sub clusters of low risk, medium risk and high risk. On the basis of attribute values of low risk, high risk and medium risk clusters compute impact of a attributes to other attribute and assign weights to different range of attribute value and this calculated weights are stored in database. When new patient records come for prediction, predict class label of that patient. It identify weather patient belongs from low risk, medium risk are high risk of cancer. Proposed approach also determines the relative pattern of attributes of cancer disease patients [5].

So this research work is fruitful for medical experts as well as patient of cancer. This research work is helpful in detection of cancer disease in early stage, better diagnosis, and provides preventive measure before the situation becomes critical. This research work uses data mining approach and beneficial for cancer patient medical experts of society. Motivation towards this project is increasing the cases of cancer in India as well as word wide. Following figure gives current status of cancer in India [6].

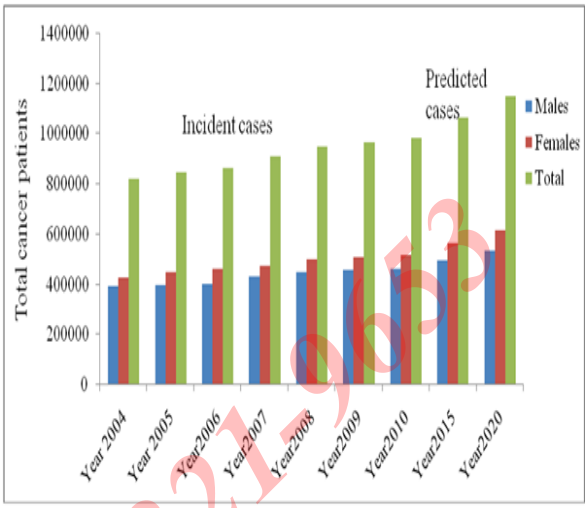


Figure 2: Year wise Incident cases and predicted cases of cancer

- Cancer is second leading disease causes death in India as well as worldwide. Total7.6 million persons are estimated death causes by cancer in year 2013.
- It is estimated that 13.7 of total deaths caused due to cancer disease in world wide.
- According to WHO reports total death in INDIA in year 2013caused by cancer is 0.556 million peoples.
- 71 % of Indian Peoples lies in 30-69 age groups. Following Table .1shows total death and death due to cancer in India in 30-69 age groups.

TOTAL DEATH (IN MILLIONS)	SEX	% DEATH DUE TO CANCER
2.5	MALE	8
1.6	FEMALE	12

ISSN: 2321-9653

IV. METHODOLOGY

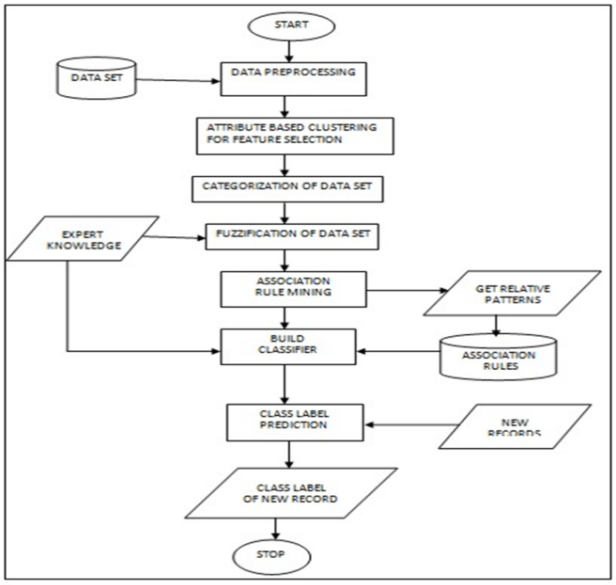


Figure 3: Flow diagram of Proposed Approach

Data Collection: Data set is taken from NANAVATI HOSPITAL mumbai. This data set contains total 11 attributes and 153 cancer patients' details. Records of some patients are incomplete and some attributes are irrelevant[7].

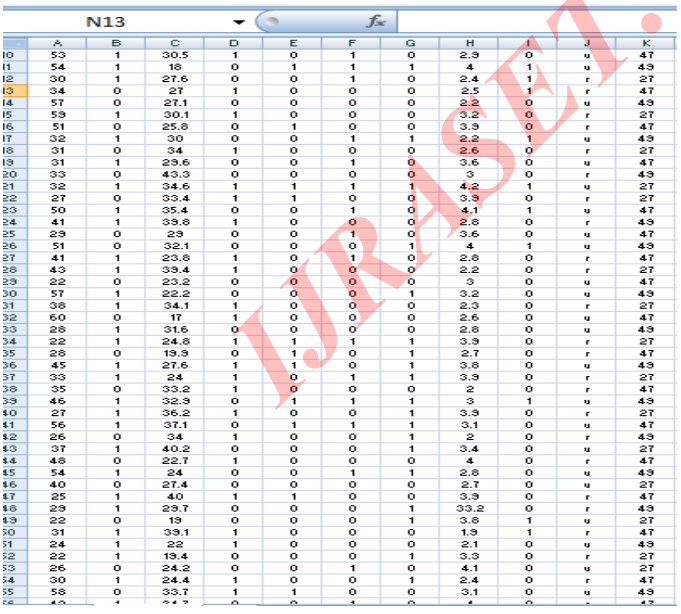


Figure4: nanavati hospital, mumbai data set key attributes of cancer disease of nanavati hospital mumbai dataset:

SNO	Attribute Name	Description
1.	Age	Age of person
2.	Sex	Whether Person is male or female
3.	BMI	Body Mass Index of the patient
4.	FamilyHistory	Whether Patient has family history of cancer or not
5.	Tuberculosis	Whether Patients has Tuberculosis or not
6.	SmokingHabit	Whether Patient has smoking habit or not
7.	LymphNodeInvolvement	Lymph node involved or not
8.	FarthestTumorSize	Farthest Tumor Size of the patient
9.	EnvironmentalEffects	Weather patients Belongs to an area having Radon, Radiation or asbestas
10.	Location	Weather patients are urban or rural
11.	Place	From which district patient belongs

Table2.NANAVATI HOSPITAL dataset contains following attributes and we get following information from expert knowledge of NANAVATI HOSPITAL, Mumbai [8]

Age: Age has been categorized into following dimensions:
0-35, 35-44, 45-54, 55-64, 65-more to show the impact of different age groups on different level of disease

SEX: IT IS CATEGORIZED AS MALE AND FEMALE.

BMI:IT HAS BEEN CATEGORIZED INTO FOLLOWING LEVELS
<15, 15 TO 18.5, 18.5 TO 25, 25 TO 30, 30 AND MORE.

Family History: Classified in to two category of patients having family history of cancer or not. Breast cancer, Colon cancer, Ovarian and some more cancer influenced because of family history.

Tuberculosis: When person having tuberculosis chance of cancer is high. This field is classified into two category whether patients having tuberculosis or not [9].

Smoking Habit: Patients are categorized in to two category patients having smoking habit or not.

ISSN: 2321-9653

Lymph Node Involvement: Patients are categorized in to two category lymph node involved or not.
Farthest Extension of Tumor: Farthest extension is classified into following categories.
<1, 1 to 2.5, 2.5 to 3.5, 3.5 and more
Environmental Effects: Classified in to two category whether Person belongs from an area having abestas, radon or radiation in the environment or not. For cancer disease environmental factor play a important role [10].
Location: Patients are categorized in to rural and urban reasons. The idea is based on available medical facility and lifestyle which in terms influence the disease.
Place: Three district codes 28, 47 and 49 would be used. It means different environment plays different pattern of diseases [11].

DATA PREPROCESSING

Missing value Handling and Data Transformation
In data preprocessing to handle missing attribute value of data set when any record of patient having more than two attributes missing we apply list wise deletion otherwise we apply Mean substitution approach. Overall data preprocessing approach includes following steps [12,13].
Step 1: Delete all the rows from data set having more than two missing attribute value.
Step 2: In remaining data set replace the missing value by substituting a mean value.
Step 3: Converting Family History, Smoking Habit, Lymph Node Involvement, Environmental Effects, Sex and Location attribute value into 0 and 1.
Step 4: Replacing the age value as shown in following table.
Step 4: Replacing the age value as shown in following table.
Fuzzificztion of attribute value and association rule mining:

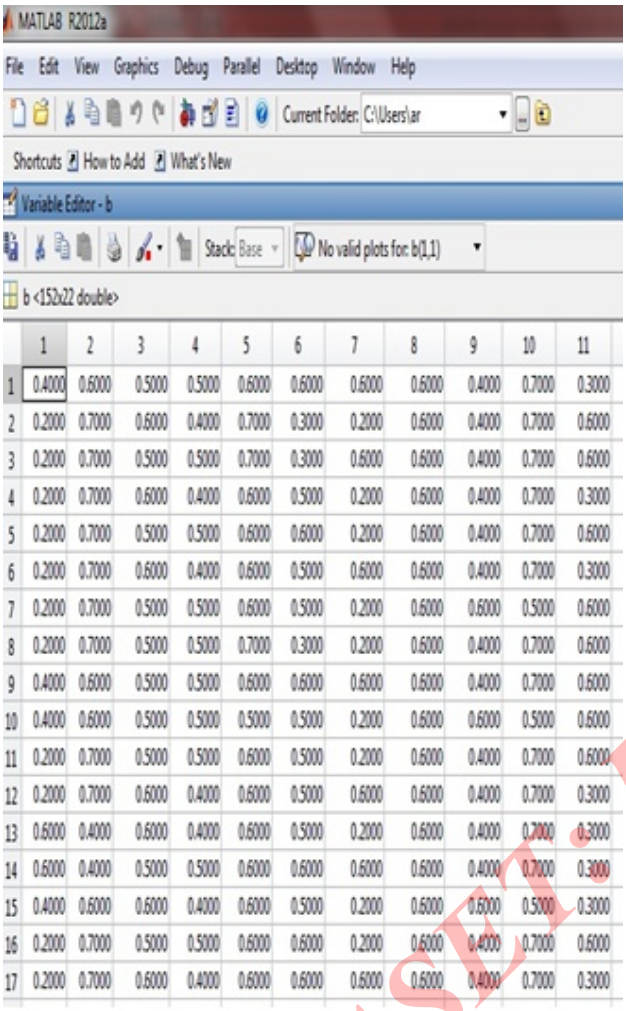
Then replace the attribute value of data set by its membership value and nonmember ship value. Membership value$\Phi(x)$ and non membership value$\Omega(x)$ for certain range of attribute value is achieved from expert's knowledge. Following table represents membership value and non membership value corresponding to attribute value [14,15].

Attribute name	0		1		2		3		4	
	Φ(x)	? (x)	Φ(x)	? (x)	Φ(x)	? (x)	Φ(x)	? (x)	Φ(x)	? (x)
Age	0.2	0.7	0.3	0.6	0.4	0.6	0.6	0.4	0.7	0.4
Family History	0.2	0.6	0.6	0.6						
Smoking Habit	0.3	0.7	0.6	0.4						
Lymph Node Involvement	0.4	0.6	0.7	0.3						
Environmental Effects	0.2	0.6	0.7	0.4						
Sex	0.6	0.4	0.5	0.5						
Location	0.4	0.6	0.5	0.6						
Farthest Extension of Tumor	0.4	0.6	0.6	0.5	0.7	0.4	0.9	0.2		
BMI	0.4	0.5	0.5	0.5	0.7	0.3	0.6	0.5	0.6	0.6
Place	0.4	0.6	0.3	0.7	0.3	0.8				
Tuberculosis	0.4	0.7	0.6	0.5						

Table3: Fuzzy value use to replace attribute values

Expert's knowledge is needed to replace the attribute value by its fuzzy value After fuzzification of attribute value we apply α-cut for α=0.6 in data set and replace all the attribute value more than or equal to 0.6 by 1 and remaining attributes value by 0[16]. Then computing number of 1 appearing for each attributes this will be the appearing time of each attributes. Using appearing time of each attributes constructing transaction table then applying apriori algorithm to determine the interdependency between attributes of cancer disease. Following steps are performed to compute association rules[1,3,7].
Step 1: We transformed the patient records using expert knowledge shown in above table.
Step 2: After applying α-cut for α=0.6 and representing useful value by 1 and remaining value by 0.
Step 3: Selecting only those attributes whose value is 1to get transaction table.
Step 4: Computing appearing time of each attribute.
Step 4: For minimum support count 4 and minimum threshold 50 % and using apriori algorithm. Algorithm used for apriori is given below.

V. EXPERIMENTS AND RESULTS



	1	2	3	4	5	6	7	8	9	10	11
1	0.4000	0.6000	0.5000	0.5000	0.6000	0.6000	0.6000	0.6000	0.4000	0.7000	0.3000
2	0.2000	0.7000	0.6000	0.4000	0.7000	0.3000	0.2000	0.6000	0.4000	0.7000	0.6000
3	0.2000	0.7000	0.5000	0.5000	0.7000	0.3000	0.6000	0.6000	0.4000	0.7000	0.6000
4	0.2000	0.7000	0.6000	0.4000	0.6000	0.5000	0.2000	0.6000	0.4000	0.7000	0.3000
5	0.2000	0.7000	0.5000	0.5000	0.6000	0.6000	0.2000	0.6000	0.4000	0.7000	0.6000
6	0.2000	0.7000	0.6000	0.4000	0.6000	0.5000	0.6000	0.6000	0.4000	0.7000	0.3000
7	0.2000	0.7000	0.5000	0.5000	0.6000	0.5000	0.2000	0.6000	0.6000	0.5000	0.6000
8	0.2000	0.7000	0.5000	0.5000	0.7000	0.3000	0.2000	0.6000	0.4000	0.7000	0.6000
9	0.4000	0.6000	0.5000	0.5000	0.6000	0.6000	0.6000	0.6000	0.4000	0.7000	0.6000
10	0.4000	0.6000	0.5000	0.5000	0.5000	0.5000	0.2000	0.6000	0.6000	0.5000	0.6000
11	0.2000	0.7000	0.5000	0.5000	0.6000	0.5000	0.2000	0.6000	0.4000	0.7000	0.6000
12	0.2000	0.7000	0.6000	0.4000	0.6000	0.5000	0.6000	0.6000	0.4000	0.7000	0.3000
13	0.6000	0.4000	0.6000	0.4000	0.6000	0.5000	0.2000	0.6000	0.4000	0.7000	0.3000
14	0.6000	0.4000	0.5000	0.5000	0.6000	0.6000	0.6000	0.6000	0.4000	0.7000	0.3000
15	0.4000	0.6000	0.6000	0.4000	0.6000	0.5000	0.2000	0.6000	0.6000	0.5000	0.3000
16	0.2000	0.7000	0.5000	0.5000	0.6000	0.6000	0.2000	0.6000	0.4000	0.7000	0.6000
17	0.2000	0.7000	0.6000	0.4000	0.6000	0.6000	0.6000	0.6000	0.4000	0.7000	0.3000

Figure6.: Result of represents data set after transforming by institionistic fuzzy value.

VI. CONCLUSION AND DISCUSSION

This study is provable and valuable to determine the hidden relative patterns of cancer disease. Feature selection is needed because in general medical data are contains lots of irrelevant attribute in proposed approach clustering technique is used for grouping the attributes in relevant and irrelevant clusters then further processing is done only on the relevant attributes data set[1,6].

Clustering approach is applied to grouping the cancer patients in to low, medium and high category. Since nature of disease are in general in fuzzy nature so applying Fuzzy based classifier are suitable for our projects. To predict relative patterns between the attributes dependency and independency applying Intuitionist Fuzzy approach is suitable and for applying Intuitionistic Fuzzy approach medical expert's knowledge is needed[4,5].

This approach uses Attribute based clustering for Feature selection, Fuzzy set, Cluster Energy, Point Energy and association rule mining to determine the association between the attribute of cancer this approach are also be applicable to determine the hidden patterns of other disease Proposed approach is suitable for determining relative patterns of cancer disease and prediction of class label of new patients. This approach is also suitable for identifying relative patterns between attributes of some more diseases. This approach is not limited till disease we can apply this approach to some other fields in long run since. This approach is helpful for medical expert's patient and emphasis on detection of cancer in early stage so fruitful for society.

VII. FUTURE WORK

Our future goal is to apply this approach on other high dimension dataset and apply this approach to other disease like heart disease, diabetes and much other disease to identify relative patterns between these diseases and predict the class label of new records for these diseases.

REFERENCES

[1] Maryvonne Miquel And Anne Tchounikine, "Software Components Integration In Medical Data Warehouses:A Proposal," Proceedings Of The 15 Th IEEE Symposium On Computer-Based Medical Systems (CBMS 2002) 1063-7125/02 $17.00 © 2002 IEEE.

[2] Daniel Ramot, Menahem Friedman, Gideon Langholz, And Abraham Kandel, "Complex Fuzzy Logic," 1063-6706/03$17.00 © 2003 IEEE.

[3] Pasiluukka And Tapiolepp¨Alampi, "Similarity Classifier With Generalized Mean Applied To Medical Data Using Different Preprocessing Methods" , 0-7803-9158-6/05/$20.00 © 2005 IEEE.

[4] Mykolapechenizkiy, Alexey Tsymbal And Seppopuuronen, "Local Dimensionality Reduction Within Natural Clusters For Medical Data Analysis," Proceedings Of The 18th IEEE

Symposium On Computer-Based Medical Systems (CBMS'05) 1063-7125/05 $20.00 © 2005 IEEE.

[5] Shoji Hirano And Shusakutsumoto, "Strucutural Comparison And Cluster Analysis Of Time-Series Medical Data".

[6] S. Cavuto And E.Grossi, "The Fuzzy Nature Of Health And Disease," 1 -4244-0363-4/06/$20.00 ©2006 IEEE.

[7] Nikhil R. Pal, "A Fuzzy Rule Based Approach To Identify Biomarkers For Diagnostic Classification Of Cancers," 1-4244-1210-2/07/$25.00 C 2007 IEEE.

[8] Mila Kwiatkowska, M. Stella Atkins, Najib T. Ayas, And C. Frank Ryan, "Knowledge-Based Data Analysis: First Step Toward The Creation Of Clinical Prediction Rules Using A New Typicality Measure," IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 11, NO. 6, NOVEMBER 2007.

[9] Sang C. Suh, Sam. Saffer And Naveen Kumar Adla, "Extraction Of Meaningful Rules In A Medical Database," 2008 Seventh International Conference On Machine Learning And Applications.

[10] Umair Abdullah, Jamil Ahmad And Aftab Ahmed, "Analysis Of Effectiveness Of Apriori Algorithm Inmedical Billing Data Mining," 2008 International Conference On Emerging Technologies IEEE-ICET 2008 Rawalpindi, Pakistan, 18-19 October, 2008.

[11] Weidong Mao And Jinghe Mao, "The Application Of Apriori-Gen Algorithm In The Association Study In Type 2 Diabetes," 978-1-4244-2902-8/09/$25.00 ©2009 IEEE.

[12] Ping-Hung Tang And Ming-Hseng Tseng , "Medical Data Mining Using Bga And Rga For Weighting Of Features In Fuzzy K-Nn Classification," Proceedings Of The Eighth International Conference On Machine Learning And Cybernetics, Baoding, 12-15 July 2009.

[13] Thannobaribarg, Siripornsupratid And Chidchanoklursinsap, "Contemporary Classification On Medical Data Based On Non-Linear Feature Extraction," 2009 International Conference On Computational Science And Its Applications.

[14] Alaqabaja, Mohammed Alshalalfa, Redaalhajjand Jon Okne, "Multiagent Approach For Identifying Cancer Biomarkers," 2009 IEEE International Conference On Bioinformatics And Biomedicine.

[15] Jan E. Szulejko, Michael Mcculloch, Jennifer Jackson, Dwight L. Mckee, Jim C. Walker, And Touradjsolouki, "Evidence For Cancer Biomarkers In Exhaled Breath," IEEE SENSORS JOURNAL, VOL. 10, NO. 1, JANUARY 2010.

[16] Azrashamim, Maqbooldin Shiek And Saifur Rehman Malik, "Intelligent Data Mining In Autonomous Heterogeneous Bio Databases," 2010 Second International Conference On Computer Engineering And Applications