Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis of Treatment of Prostate Cancer by Using Multiple Techniques of Data Mining

Nazreena Rahman (MTech Student)*
Department of Information Technology
Institute of Science and Technology
Gauhati University
Guwahati, India

Parismita Sarma(Assistant Professor)
Department of Information Technology
Institute of Science and Technology
Gauhati University
Guwahati, India

Abstract- The Healthcare industry is the most significant amongst the information intensive industries. Medical information, knowledge and data keep growing on a daily basis. Therefore, huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. This leads to the use of data mining in medical informatics. This paper attempts to make the efficient use of a database "Prostate Cancer Dataset" and try to perform an integration of modified clustering and classification techniques of data mining to find the stages and treatments particularly for the prostate cancer disease. The result of the experiment shows that the modified technique gives more promising and reliable results with utmost accuracy.

Keywords-: Data Mining, modified clustering, classification, Prostate Cancer Dataset

I. Introduction

Because of the complexity of some medical cases and diseases, previous analyzing methods have not been able to fulfill both patients' and researchers' needs and expectations. Therefore, it is essential to choose a unique, precise, and efficient analyzing method through which the existing data and medical records of patients can be analyzed carefully and precisely. Meaningful analysis of medical records provides the researchers with specific, new and useful clinical knowledge and understanding of the patients' health. At present period, Cancer is progressively reaching epidemic proportions. Number of cancer deaths has been increasing each year. Along with various types of cancers, prostate cancer is the most common cancer amongst men. Prostate cancer remains one of the leading causes of cancer death with a reported incidence rate of 650,000 cases per annum worldwide [1].

In this report, an effort is made to find out the stages of prostate cancer patients efficiently in order to give them the best treatment. In Cancer Health Care, Data Mining is becoming increasingly popular therefore, many data mining techniques are applied efficiently in cancer research [2]. Amongst them, classification is one of the forms of data analysis that can be used to extract models describing important data classes or to predict categorical labels in data mining (DM). Classification can be applied to the medical databases which will classify data with a reasonable accuracy. It is a supervised learning problem of assigning an object to one of the several predefined categories based upon the attributes of the object. It is an important problem studied extensively in data mining and machine learning. Among the numerous classifiers in existence, it is found that decision tree classifiers are relatively efficient, correct and accurate compared with the others and they are more interpretable [3] [4].

Today, C4.5 has probably become the most widely used and studied decision tree construction algorithm [5]. It induces decision trees and rules from datasets, which could contain categorical and numerical attributes. Rules can readily be expressed in a language that human can understand them, or in a database access language, such as SQL. Though C4.5 is a well known algorithm used for classifying dataset but when it is used in mass calculations, the efficiency has been found low. However, to make the algorithm more practically to the dataset, another unsupervised data mining technique k-means clustering is introduced here to integrate with C4.5 algorithm in a modified way to increase the efficiency of C4.5 algorithm [3] [4] [6].

II.Overview Of Prostate Cancer

The prostate is a gland in the male reproductive system that produces the majority of seminal fluid that carries sperm. The walnut-sized gland is located beneath a man's bladder and surrounds the upper part of the urethra, the tube that carries urine from the bladder. Prostate function is regulated by testosterone, a male sex hormone produced mainly in the testicles.

Several types of cells are found in the prostate, however, it is evident that almost all prostate cancers develop from the gland cells. Gland cells make the prostate fluid that is added to the semen. The medical term for a cancer that starts in gland cells is *adenocarcinoma*. Other types of cancer can also start in the prostate gland, including sarcomas, small cell carcinomas, and transitional cell carcinomas. But these types of prostate cancer are so rare.

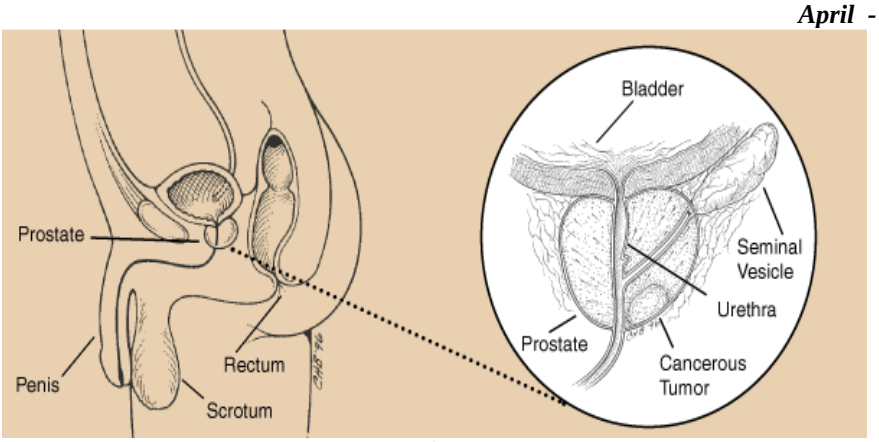


Fig. 1 Overview of the Prostate part

The stage of prostate cancer is one of the most important factors of choosing the best way for its treatment. In the beginning, prostate cancer has been diagnosed, graded, and staged .After that depending on the situation, the treatment is given to the patient .The treatment options for men with prostate cancer may include:

- Expectant management (watchful waiting) or active surveillance
- Surgery
- Different types of Radiation therapy
- Cryosurgery (cry therapy)
- Different types of Hormone therapy
- Chemotherapy
- Vaccine [7].

III. Introduction To The Dataset

Prostate cancer is the most common malignancy among elderly male. Nearly two thirds are diagnosed in men aged 65 or older and it is rare before age 40. The average age at the time of diagnosis is about 67. Therefore, it is very much necessary to determine the stages of prostate cancer in order to give the proper treatment to the patient. Moreover, the age factor also is a significant reason which contributes much complication to the treatment. Normally, to define the stage of prostate cancer TNM system can be used.This system is one of the most widely used staging systems.TNM (using T, N and M categories) system has been accepted by the International Union Against Cancer (UICC) and the American Joint Committee on Cancer (AJCC). TNM system gives four types of stages for prostate cancer [7]. But along with the TNM system, age should also be considered as an important factor in so far as prostate cancer treatment is concerned. In this work, six parameters have been chosen to find the stages of prostate cancer. First one is Gleason Score. It is used to grade the prostate cancer disease. The value of Gleason Score is between 2 to 10. Cancers with a Gleason Score of 6 or less are often called well-differentiated or low grade. As the score increases, grade of the cancer also increases accordingly. Second one is the Age and it is found that prostate cancer begins from age 40. Third one is the PSA. PSA stands for Prostate Specific Antigen which is found in the serum of a patient and can be easily measured by a routine biochemical procedure. The PSA test is a simple blood test that detects a specific protein produced by the prostate. Usually a serum level above 4 ng/ ml is taken as an indicator of the possible presence of prostate cancer and used as the trigger for further clinical evaluation. Several studies using large number of men suggested that the quantization of serum PSA was a useful diagnostic tool for detecting the presence of prostate cancer.

Fourth attribute is the T categories (clinical), the extent of the primary tumour (T category). There are 4 categories for describing the local extent of a prostate tumour, ranging from T1 to T4.

- T1 means the doctor can not feel the tumour or see it with imaging such as transrectal ultrasound.
- T2 means doctor can feel the cancer with a digital rectal exam (DRE) or see it with imaging such as transrectal ultrasound but it still appears to be confined to the prostate gland.
- T3 stand for the cancer has begun to grow and spread outside the prostate and may have spread into the seminal vesicles.
- T4: The cancer has grown into tissues next to your prostate (other than the seminal vesicles), such as the urethral sphincter (muscle that helps control urination), the rectum, the bladder, and/or the wall of the pelvis.

Then fifth attribute comes as the N categories. N categories describe whether the cancer has spread to nearby (regional) lymph nodes.

- NX: Nearby lymph nodes were not assessed.
- N0: The cancer has not spread to any nearby lymph nodes.
- N1: The cancer has spread to one or more nearby lymph nodes in the pelvis.

M categories describe whether the cancer has spread to distant parts of the body. The most common sites of prostate cancer spread are to the bones and to distant lymph nodes, although it can also spread to other organs such as lungs and liver.

- M0: The cancer has not spread past nearby lymph nodes.
- M1: The cancer has spread beyond the nearby lymph nodes.

The study work collected the experimental dataset of prostate disease patients from Dr. B. Barooah Cancer Institute [8].Dataset reveal that there are eleven types of treatments which can be given to the prostate cancer patients.
The classes are

1. Cancer stage is first (1): Treatment is external beam radiation therapy and radical prostatectomy. (written in classification tree as firstnebrtnrp)
2. Cancer stage is first (1): Treatment is brachy therapy and radical prostatectomy. (written in classification tree as firstnbtnrp)
3. Cancer stage is first (1): Treatment is brachy therapy and chemo therapy. (written in classification tree as firstnbtnct)
4. Cancer stage is second (2): Treatment is radical prostatectomy and external beam radiation therapy.(written in classification tree as secondnrpnebrt)
5. Cancer stage is second (2): Treatment is brachy therapy and external beam radiation therapy.(written in classification tree as secondnbtnebrt)
6. Cancer stage is second (2): Treatment is external beam radiation therapy.(written in classification tree as secondnebrt)
7. Cancer stage is three (3): Treatment is external beam radiation therapy and hormonal therapy. (written in classification tree as thirdnebrtnht)
8. Cancer stage is three (3): Treatment is external beam radiation therapy and brachy therapy and short course of hormonal therapy and chemo therapy. (written in classification tree as thirdnebrtnbtnshtnct)
9. Cancer stage is four (4): Treatment is external beam radiation therapy and hormonal therapy. (written in classification tree as fourthnebrtnht)
10. Cancer stage is four (4): Treatment is external beam radiation therapy and hormonal therapy and chemo therapy. (written in classification tree as fourthnebrtnhtnct)
11. Cancer stage is four (4): Treatment is hormonal therapy and rarely chemo therapy. (written in classification tree as fourthnhtnract)

IV. Proposed Method

Here we attempt to increase the efficiency of the existing algorithm C4.5 so that it can be more practically feasible and applicable. To obtain a better-structured decision tree, we combine the classification technique with the k-means clustering in a modified and an efficient way by reducing the drawback of k-means algorithm. We find that the drawback of k-mean is the prerequisite of the earlier declaration of no of clusters. Therefore, to overcome this problem and for making right decision of taking the number of clusters efficiently we use Calinski Harabasz (CH) Index [9].
The Output of the Proposed Algorithm is as follows
First step: Applying modified k-means algorithm on those continuous attributes (Gleason Score, Age and PSA). According to the proposed algorithm, we first determine the number of clusters by using Calinski Harabasz index. Here, optimum number of clusters in a data is the value of K which maximizes the CH_K .Therefore, we find the k which maximizes CH index. Calinski Harabasz (CH) Index is as follows.

This index for N data points and N_c clusters is computed as

$$CH(N_c) = \frac{\{trace\, B/(N_c - 1)\}}{\{trace\, W/(N - N_c)\}}$$

B and W are the between and within cluster scatter matrices. The trace of the between cluster scatter matrix B can be written as

$$trace\, B = \sum_{i=1}^{N_c} N_i \left\| \mathbf{c}_i - \mathbf{c}_0 \right\|^2$$

where Ni is the numbers of points in cluster i, and c_o is the centroid of the entire dataset. The trace of the within-cluster scatter matrix W can be written as

$$trace\, W = \sum_{j=1}^{N_c} \sum_{i=1}^{N_i} \left\| \mathbf{x}_i - \mathbf{c}_j \right\|^2$$

Therefore, the CH index can be written as

$$CH(N_c) = \left\{ \frac{\sum_{i=1}^{N_c} N_i \left\| \mathbf{c}_k - \mathbf{c}_0 \right\|^2}{N_c - 1} \right\} / \left\{ \frac{\sum_{i=1}^{N_c} \sum_{j=1}^{N_i} \left\| \mathbf{x}_j - \mathbf{c}_i \right\|^2}{N - N_c} \right\}$$

Second Step: After finding the number of clusters we apply the k-mean algorithm and find the mean values of the clusters.

Third Step: Assign the each set of Gleason Score, Age and PSA values to the clusters. Highest priority is given to the Gleason Score as this value plays the vital role to find out the appropriate treatment of prostate cancer. After Gleason Score, priority is given to the Age value because elderly people can not take all kinds of treatments due to their age. PSA value is given priority at length.

Fourth Step: Apply C4.5 algorithm on the clustered data and other categorical data.

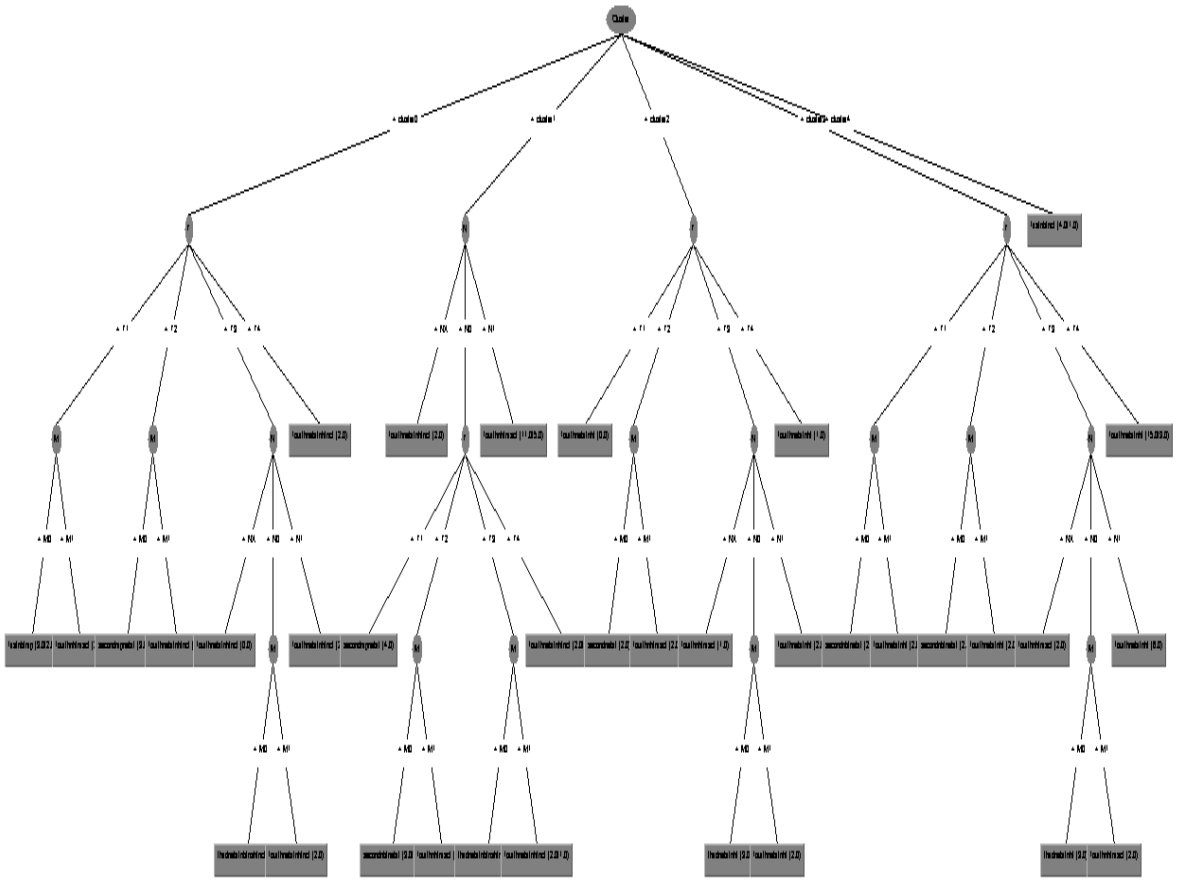
V. Experminental Results

Here, the C4.5 algorithm is verified by the analysis of prostate cancer dataset. We implement the proposed algorithm in the WEKA data mining tool [10]. In order to perform experiment using system WEKA, the file format for prostate cancer database has been changed to .arff.

First apply Calinski Harabasz (CH) Index on the dataset and try to find the number of clusters .At this point, after applying Calinski Harabasz (CH) Index we get five numbers of clusters. Subsequent to that, we get mean values for Gleason Score, Age and PSA. The mean values for each cluster are

Cluster	Gleason Score	Age	PSA
Cluster 0	6	66	5.91
Cluster 1	7	69	8.70
Cluster 2	9	69	6.60
Cluster 3	9	70	12.8
Cluster 4	4	51	4.76

Later,the set of Gleason Score,Age and PSA values are assigned to the each cluster according to the priority. Finally, C4.5 algorithm is applied on the clustered data and other categorical data (T,N and M) to get the final result.




```
Cluster = cluster0
|   T = T1
|   |   M = M0: firstnbtnrp (3.0/2.0)
|   |   M = M1: fourthnhtnract (2.0)
|   T = T2
|   |   M = M0: secondnrpnebrt (3.0/1.0)
|   |   M = M1: fourthnebrtnhtnct (2.0)
|   T = T3
|   |   N = NX: fourthnebrtnhtnct (0.0)
|   |   N = N0
|   |   |   M = M0: thirdnebrtnbtnshtnct (2.0)
|   |   |   M = M1: fourthnebrtnhtnct (2.0)
|   |   N = N1: fourthnebrtnhtnct (2.0)
|   T = T4: fourthnebrtnhtnct (2.0)
Cluster = cluster1
|   N = NX: fourthnebrtnhtnct (2.0)
|   N = N0
|   |   T = T1: secondnrpnebrt (4.0)
|   |   T = T2
|   |   |   M = M0: secondnbtnebrt (3.0/1.0)
|   |   |   M = M1: fourthnhtnract (2.0)
|   |   T = T3
|   |   |   M = M0: thirdnebrtnbtnshtnct (3.0)
|   |   |   M = M1: fourthnebrtnhtnct (2.0/1.0)
|   |   T = T4: fourthnebrtnhtnct (2.0/1.0)
|   N = N1: fourthnhtnract (11.0/5.0)
Cluster = cluster2
|   T = T1: fourthnebrtnht (0.0)
|   T = T2
|   |   M = M0: secondnebrt (2.0)
|   |   M = M1: fourthnhtnract (2.0)
|   T = T3
|   |   N = NX: fourthnhtnract (1.0)
|   |   N = N0
|   |   |   M = M0: thirdnebrtnht (3.0)
|   |   |   M = M1: fourthnebrtnht (2.0)
|   |   N = N1: fourthnebrtnht (2.0)
|   T = T4: fourthnebrtnht (1.0)
Cluster = cluster3
|   T = T1
|   |   M = M0: secondnbtnebrt (2.0)
|   |   M = M1: fourthnebrtnht (2.0)
|   T = T2
|   |   M = M0: secondnbtnebrt (2.0)
|   |   M = M1: fourthnebrtnht (2.0)
|   T = T3
|   |   N = NX: fourthnhtnract (2.0)
|   |   N = N0
|   |   |   M = M0: thirdnebrtnht (3.0)
|   |   |   M = M1: fourthnhtnract (2.0)
|   |   N = N1: fourthnebrtnht (6.0)
|   T = T4: fourthnebrtnht (15.0/3.0)
Cluster = cluster4: firstnbtnct (4.0/1.0)
```

Fig. 2 Decision tree and Rules during the proposed classification of Prostate Cancer Dataset

VI. Conclusions

In this work, the C4.5 algorithm is verified by the analysis of prostate cancer dataset. However, with the help of the proposed algorithm, it is found that it increases the important attribute's influence. The findings of the present study suggest that the treatment of the prostate cancer patient's depends also on Age attribute. Age should also be taken into account when looking into the treatment options. Age plays a crucial role to determine the proper treatment. On the basis of the different ages of the patients, different treatments can be given to them though their stages of the disease are same. Therefore, we get more consistent and accurate rules. Moreover, it is also evident that decision rules are very useful in classifying unknown dataset for the biologists, clinicians and for oncologists. In future, there is a possibility to combine more data mining techniques by integrating classification, clustering and association techniques of data mining to get more faster algorithm. By doing this, it may be expected that faster and more effective results can be achieved. As the incidences of the prostate cancer are rising year by year, therefore, the research work on this particular cancer disease is

very important. It does not only support the knowledge of the data mining techniques and medical research field in prostate cancer treatment, but some new conclusions can also be put forward that can prove beneficial for further research on prostate cancer.

Refernces

[1] Balkrishna B Yeole, "Trends in the Prostate Cancer Incidence in India", *Asian Pacific Journal of Cancer Prevention*, Vol9, pp.141-144, 2008.

[2] Kemal Hakan GULKESEN, Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University , Antalya – TURKEY and Department of Health Informatics, Informatics Institute, Middle East Technical University, Ankara–TURKEY, İsmail Turker KOKSAL, Department of Urology, Faculty of Medicine, Akdeniz University, Antalya – TURKEY, Sebahat OZDEM, Central Laboratory Clinical Biochemistry Unit, Akdeniz University Hospital, Antalya – TURKEY, Osman SAKA, Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University , Antalya – TURKEY, "Prediction of prostate cancer using decision tree algorithm",pp.681-686,2010.

[3] Arun K.Pujari, "Data Mining Technology", *Universities Press*, Second Edition, 2011.

[4] Jiawei Han and Micheline Kamber,"Data Mining Concepts and Techniques", *Morgann Kaufmann*, Second Edition, 2006.

[5] Zhu Xiaoliang ,Computing Center, Wang Jian,College of Science dept, YanHongcan College of Science dept, Wu Shangzhuo College of Chemical Engineering and Biological Technology, Hebei Polytechnic University, China "Research and Application of the improved Algorithm C4.5 on Decision Tree" *International Conference on Test and Measurement*,pp.184-187, 2009 .

[6] Varun Kumar, Proffesor &Head, Department of Computer Science and Engineering ,ITM University Gurgaon, India ,Nisha Rathee ,persuing M.Tech,Department of Computer Science and Engineering ,ITM University, "Knowledge discovery from database Using an integration of clustering and classification", *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 2, No.3,pp.29-33, March 2011.

[7] (April 06, 2013)American Cancer Society, "All about Prostate Cancer Overview". [Online]. Available: www.cancer.org/

[8] (September 19, 2012) Dr. B. Borooah Cancer Institute. [Online]. Available: http://www.bbcionline.org/

[9] Hamidreza Bayati, Heydar Davoudi, Emad Fatemizadeh, "A Heuristic Method for Finding the Optimal Number of Clusterswith Application in Medical Data*",* Department of Electrical Engineering,Sharif University of Technology , Vancouver, British Columbia, Canada,*30th Annual International IEEE EMBS Conference*,pp.4684-4687, August 20-24,2008.

[10] Weka 3- Data Mining with open source machine learning software available from: - http://www.cs.waikato. ac.nz/ml/ weka/

