International Journal of
Applied Mathematics,
Electronics and Computers
ISSN: 2147-8228 http://ijamec.atscience.org

Advanced Technology and Science

Original Research Paper

Decision Tree Application for Renal Calculi Diagnosis

Murat TOPALOĞLU*[1], Gözde MALKOÇ[1]

Accepted 3rd September 2016

Abstract: Data mining is used for the extraction of secret, valuable and usable data from the big data and to provide strategic decision support. It created a new perspective for the use of the data in healthcare in addition to finding the answers of unexplored questions. It has gained wider usage as a method. The aim of this study is to develop a decision tree and a list of rules by data mining for the early diagnosis of renal calculi. A data set including blind and retrospective data for 150 people can diagnose with 6 attributes. A decision support system analysis was developed for the diagnosis of the patients with suspected renal calculi. Based on the results obtained and the analysis developed, a decision tree and list of rules were created to determine the factors that affect renal calculi. Weka program and J48 algorithm were used to create the decision tree and the list of rules and it was found to be 74.63% successful.

Keywords: *Data Mining, Decision Tree, Renal Calculi Diagnosis, J48.*

1. Introduction

The aim of medical informatics is to use computer and communication technologies that are interacting with other sciences in order to utilize, analyse and reconstruct medical information in an effective way. Medical informatics helps to obtain valid, detailed and reliable results on a global scale as it exponentially improves the data collection, process and evaluation capacities in medical centers [1].

A renal calculus is among the most common diseases in Turkey. About 15% of women and 5% of men in Turkey have been diagnosed with nephrolithiasis. It generally occurs in thirties for the first time. Kidneys are like filters in the human body and they help the disposal of waste through the urinary tract. Sometimes some of this waste might cumulate in the kidneys. Renal calculi are the solid pieces that are formed with the crystallization of calcium, phosphate and other minerals in the urine.

Drinking inadequate amount of water, obesity, consuming too much sugar and salt might cause nephrolithiasis as well as genetic factors. Most of the renal calculi are calcium stones resulting from the accumulation of the calcium in the kidneys.

In addition to the environmental factors in the formation of renal calculi, genetic factors may be the primary cause. Not having enough water is one of the biggest risk factors for renal calculi. It is often recommended to drink about 2 liters of water daily. Besides, it is good for our well-being in general not to consume too much sugar and salt, to follow a balanced diet, to stay away from convenience food and not to be overweight. Renal calculi may also stay in the kidneys without showing any symptoms or doing any harm to the kidneys. Following the general information given for renal calculi, we are going to talk about data mining.

Data mining is the extraction of secret, valuable and usable information from a big amount of data. Data mining which is used to provide strategical decision support aims to construct decision making models based on analysis methods.

It creates directive models for decision making techniques for medical institutions where the data is collected for analyses [2]. It is important to emphasize the difference between flaws and misappropriations in health services, to minimize the risks and to take necessary precautions based on this distinction for patients' security and well-being [3] [4]. A large amount of information can be retrieved from the various data sets used in daily life [5, 6]. Data mining has made life easier by enabling us to access to more data in a shorter time span [7]. In the simplest term, data mining is the extraction of imprecise, valid and applicable data from data stacks with a dynamic process [8, 9]. It is possible to obtain valuable information from the big data stacks using data mining and statistical analysis techniques. This information helps doctors during the decision process with computer-aided diagnosis work and contributes to the improvement of health care applications [10].

Although potent devices have been produced by expert systems, they have not gained common use because of the field-related data changing rapidly and the diversity of views among experts [11]. Data warehouse is a large store of data accumulated from a wide range of sources under the same roof [12]. The steps in data mining include the definition of the problem to be solved, the obtaining of previous data about the problem, the selection of usable data, housekeeping of the data, the evaluation of analysis results, and the utilization of these results [13, 14]. The time saving and accelerating effect of the data selection will be obvious in the later stages [15].

A decision tree is an approach commonly used in data mining for categorization and estimations. Decision trees benefit the decision makers as they are easy to interpret and understand [16, 17].

1.1. The C5.0 Algorithm

It is one of the most common decision tree algorithms, especially used for big data sets. C5.0 helps us to get more proper decision trees in terms of form [18]. WEKA is an open coded data mining program with a functional graphics interface which keeps machine learning algorithms together [19, 20].

2. Recent Studies

İlkuçar examined the chronic kidney disease with the help of artificial neural networks in his study. He emphasized what is to be done and which tests to be run for early diagnosis of chronic kidney disease [21]. Danacı et al. focused on the diagnosis of breast cancer cells using data mining methods [22]. In the study conducted by Yurtay et al. they did a data mining research for

[1] *Department of Computer Technology & Information System, School of Applied Sciences, Trakya University, Kesan Yusuf Çapraz Hersekzade Campus, 22800, Keşan/Edirne/Turkey*
** Corresponding Author: Email: murattopaloglu@trakya.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

anaemia diagnosis. In a study on iron deficiency anaemia, the system is run with decision trees [23]. Özkan et al. tried to improve the diagnosis accuracy of the laboratory tests used for the diagnosis of fibromyalgia syndrome which are supported by sympathetic dermal response parameters. SSR parameters and laboratory tests which were calculated by Matlab were analyzed with artificial neural networks and the percentages of accuracy were found [24]. Kökver et al. probed the factors affecting hypertension with data mining methods. They developed a diagnosis system that will estimate whether the patients have hypertension or not [25].

Kusiak et al. work through a decision support system which will determine whether the lung tumor is benign or malignant [26]. In their research, Topaloğlu & Sur created a decision support system to diagnose hepatitis and to minimize the number of wrong diagnoses. It will help the doctors with the diagnosis of hepatitis [27].

3. Recent Studies

3.1. The Aim of the Study

The aim of the study is to make the diagnosis of renal calculi easier and help doctors with the process. A decision tree and list of rules were created based on the full urinary tests for the diagnosis of the disease. It is possible to avoid misdiagnoses with the decision tree and the list of rules based on facts. Patients can benefit from early diagnosis of a disease they have.

Weka program was utilized for the design of the list of rules and the decision tree. The values of the patients were categorized and the roots and the branches of the decision tree were determined.

3.2. The Importance of the Study

Urinary analysis is one of the most common methods used for the diagnosis of renal calculi. Especially substances like uric acid and calcium present in normal urine become crystallized and form the structures called renal calculi. These formations can have great negative effects on a patient's daily life. The diagnosis of the disease of which treatment process is quite hard is also of utmost importance.

Data Set

For this study, a data set including retrospective and blind data from 150 people. Six attributes were used for the diagnosis of the disease. The seventh attribute is the comments "yes" and "no" for the diagnosis. The six attributes used are as follows;

- Leucocyte,
- Urine color,
- PH,
- Bilirubin,
- Appearance,
- Erythrocyte.

3.3. The Decision Tree Model and Algorithm to Be Applied to the Data Set

A decision tree and a list of rules were formed with the Weka program used in the study. Before that, the data set was processed in csv format and the data was converted to arff format and transferred to Weka Program. J48 algorithm was used to determine the factors contributing to the formation of renal calculi and for the diagnosis. RISK was selected as the root node. The results were transferred with the decision tree and the list of rules.

4. Findings

Full urine analysis can provide us with very important information of the presence of kidney problems and activities. After the full urine analyses that make up the data set were categorized, they were transferred to Excel program. The file was saved in csv format for WEKA program.

4.1. The Formation of the Decision Tree with Weka

J48, which is a decision tree algorithm, was used in the program. The last one of the training set with seven attributes is the "RISK" attribute where the diagnosis is made. The others are as follows; Leucocyte (LOW, MILD, HIGH), Urine Color (COLORLESS, STRAW-COLORED, YELLOW), PH (LOW, NORMAL, HIGH), Bilirubin (NEGATIVE, POZITIVE), Appearance (CLEAR, BLURRED, SLIGHTLYBLURRED, HEAVILYBLURRE), Erythtocyte (TRACE AMOUNT, RARE, NORMAL). 55% of the data obtained were used for training while the rest was used for testing. Classification model provides information about the structure and the size of the tree. This model belongs to the learning set.

4.2. Full Urine Analysis and Value Ranges

If Leucocyte value is between; 0 and 3 □ Low, 3 and 5 □ Mild, 5 and above □ High

If PH value is between; 4,5 and 5,5 □ Low, 5,5 and 6,5 □ Normal, 6,5 and 7,5 □ High.

If Erythrocyte value is between; 0 and 4 □ Rare, 4 and 9 □ Normal, 10 and above □ Trace Amount

The relation and attribute expressions of the arrf file is shown below;

@relation kidney disease
@attribute LEUCOCYTE string
@attribute COLOR string
@attribute PH string
@attribute BILIRUBIN string
@attribute APPEARANCE string
@attribute ERYTHROCYTE string
@attribute DISEASE string
@Data

The number of the leaves is 7 while the size of the tree is 10.

Evaluation module gives the classification error and the Kappa (=0.482) statistic both. The mean absolute error and the root mean squared error of the category probability estimations assigned by the tree are found 0.3049 and 0.4345, respectively. Classification performance was calculated as 74.63%. This value shows that the decision tree is 74.63% successful.

The algorithm excludes meaningless variables automatically and it makes the selection of the variables itself during the new learning process [28]. Here, the gender was considered to be a meaningless variable and it was excluded from the decision tree variable order. Activity was chosen as the root node. The decision tree created based on J48 algorithm is presented in Figure-1.

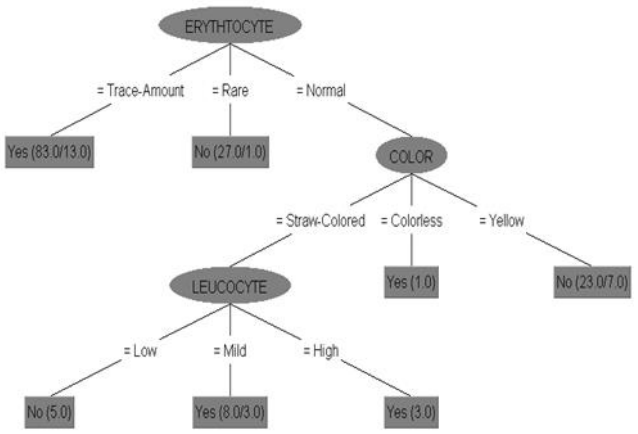

Figure 1. Renal Calculi Decision Tree

The list of rules based on the decision tree created by J48 algorithm is as follows;

1. If ERYTHROCYTE = Trace amount, then Risk = Yes
2. If ERYTHROCYTE = Rare, then Risk = No
3. If ERYTHROCYTE = Normal and Color= Colorless, then Risk = Yes
4. If ERYTHROCYTE = Normal and Color = Yellow, then Risk = No
5. If ERITROSIT = Normal and Color = Straw-colored and LEUCOCYTE = Low, then Risk = No
6. If ERITROSIT = Normal and Color = Straw-colored and LEUCOCYTE = High, then Risk = Yes
7. If ERITROSIT = Normal and Color = Straw-colored and LEUCOCYTE = Medium, then Risk = Yes

5. Conclusion

Full urine analysis values were examined with the method used in this study. It is possible to diagnose renal calculi without a surgical examination of the patient. Thus, it helps to take the necessary precautions for nephrolithiasis.

According to the design of the decision tree, the most important attribute to take into account is the amount of Erythrocyte. If the amount of the Erythrocyte is above 10, then a pathological test needs to done with ultrasonography. Again, if the amount of the Erythrocyte is between 4 and 9, then the color of the urine, which is among the attributes, should be examined. If the color is "Yellow", it means that you do not have any renal calculi. If the color is "Straw-yellow", then the amount of the Leucocyte should be examined. If the amount of Leucocyte is more than 3, then you are very likely to have renal calculi. If it is below 3, then you do not have it. If the color of the urine is "Colorless", then you need to have and pathological test. If the amount of the Erythrocyte is below 4, it means that you are less likely to have kidney problems.

The application we designed using data mining methods would help the diagnosis and treatment processes in medicine. It would be very beneficial for biological/medical fields, both clinically and executively, to get help from data mining. It is important to use the results in a safe way in every stage of diagnosis and treatment processes. The aim is to use the decision tree results through the data mining application based on the results obtained. Besides, the first step towards the formation of decision support system was taken with the list of rules created.

References

[1] E. Musoğlu, Sağlıkta Tıp Bilişiminin Önemi ve Dünyada Son Durum, Tıp Bilişimi Güz Okulu Dergisi. 2003, p. 4.

[2] Sincan M., Birinci Basamak Sağlık Hizmetleri İçin Bilişim Rehberi, Sürekli Tıp Eğitimi Dergisi. 2000.

[3] Yang W.S and Hwang S., A Process-Mining Framework for the Detection of Healthcare Fraud and Abuse, Expert Systems with Applications. 2006, vol.31, p.56-68.

[4] Can M.B., Çamur E., Koru M. and Rzayeva Z., Veri Kümelerinden Bilgi Keşki: Veri Madenciliği, Başkent Üniversitesi Tıp Fakültesi XIV. Öğrenci Sempozyumu. Ankara, 2008.

[5] Chandor A., The Penguin Dictionary of Computers, New York: Penguin Books. 1989, p.106.

[6] Albayrak M., "The detection of an epileptiform activity on EEG signals by using data mining process", PhD Thesis, Graduate School of Natural and Applied Sciences, Sakarya University, Turkey, 2008.

[7] Azimli M., "Tıpta Veri Madenciliği Uygulamaları", MSc Thesis, Institute of Information, Gazi University, Turkey, 2011.

[8] Baykal A., Application Fields of Data Mining, Dicle University Journal of Ziya Gökalp Faculty of Education, 2006, p.95-107.

[9] Zhou Z. H., Three Perspectives of Data Mining, Artificial Intelligence. Elsevier, 2003, p.139-146.

[10] Makinacı M. and Güneşer C., Göğüs Kanseri Verilerinin Sınıflandırılması, Elektrik Elektronik-Bilgisayar Mühendisliği 12. Ulusal Kongresi, 2007.

[11] Yıldırım P., Uludağ M. and Görür A., Hastane Bilgi Sistemlerinde Veri Madenciliği, Akademik Bilişim Dergisi, 2008.

[12] Shah S.C. and Kursak A., Data Mining and Genetic Algorithms Based Gene / SNP Selection, SHAH, Artificial Intelligence in Medicine. 2004, vol. 31, p.183-196.

[13] Kamrani A., Rong W. and Gonzalez R., A Genetic Algorithm Methodology for Data Mining and Intelligent Knowledge Acquisition. Computers & Industrial Engineering, 2001, vol. 40, p.361-377.

[14] Feelders A., Daniels H. and Holsheimer M., Methodological and Practical Aspects of Data Mining. Information & Management, 2000, vol.37, p. 271-281.

[15] Subramanian A., Smith L.D., Nelson A.C., Campbell J.F. and Bird D.A., Strategic Planning for Data Warehousing. Information & Management, 1997, vol.33, p.99-113.

[16] Chien C.F. and Chen L.F., Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry. Expert Systems with Applications, 2008, vol.34, p. 280-290.

[17] Özekes S. and Çamurcu Y., Veri Madenci¬liğinde Sınıflama Ve Kestirim Uygulaması, Marmara Üniversitesi Fen Bilimleri Dergisi, 2002, vol.18, p.1-17.

[18] Quinlan J.R., C4.5: Programs for Machine Learning, Elsevier, 2014.

[19] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I.H., The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 2009, vol.11, p.10-18.

[20] Witten I.A., Frank E. and Hall M.A., Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, USA, 2011.

[21] İlkuçar M., Diagnosis Chronic Kidney Disesa with Artificial Neural Network and Radial Basis Function Network, The Journal of Graduate School of Natural and

Applied Sciences of Mehmet Akif Ersoy University. 2015, vol.6, p.82-88.

[22] Danacı M., Çelik M. and Akkaya A.E., Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi, ASYU Conference, 2010.

[23] Yurtay Y., Salman Y. and Gençali F., Kansızlık Tanısına İlişkin Bir Veri Madenciliği Uygulaması, ISITES 2013, p.896-900.

[24] Özkan Ö., Yıldız M. and Köklükaya E., Improving Diagnostic Accuracy by Supporting The Laboratory Tests Which Used for Diagnosıs of Fibromyalgia Syndrome With The Sympathetic Skin Response Parameters, Sakarya University Journal of Science, 2011, vol.15, p.1-7.

[25] Kökver Y., Barışçı N., Çiftçi A. and Ekmekçi Y., Determining Affecting Factors of Hypertension with Data Mining Techniques, NWSA Academic Journal. 2014, vol.9, p.15-25.

[26] Kusiak A., Kernstine K.H., Kern J.A., McLaughlin K.A. and Tseng T.L., Medical and Engineering Case Studies, 2000.

[27] Topaloğlu M. and Sur H., Decision Tree Application to Reduce Incorrect Diagnosis in Symptoms of Jaundice, Nobel Medicus. 2015, vol.11, p.64-73.

[28] Çakır M., Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makine Öğrenmesi Teknikleri: Amprik Uygulamalar ve Karşılaştırılmalı Analiz. Uzmanlık Yeterlilik Tezi, Ankara, T.C. Merkez Bankası, 2005.