Diagnosis of Irritable Bowel Syndrome using data mining techniques

SUBMISSION DATE / POSTED DATE

22-12-2020 / 28-12-2020

CITATION

Elmi, Yasser; Rad, Arezoo Elhami (2020): Diagnosis of Irritable Bowel Syndrome using data mining techniques. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.13474656.v2

DOI

10.36227/techrxiv.13474656.v2

Diagnosis of Irritable Bowel Syndrome using data mining techniques

Arezoo Elhami Rad[a] ,Yasser Elmi[a,*]

[a,] *Department of Computer Engineering, Sabzevar Branch, Islamic Azad University, Sabzevar, Iran*

([,] Corresponding Author, E-mail: elmi@iaus.ac.ir)*

Abstract:

Chronic gastrointestinal disorders impose a heavy economic burden and stress on society and the health care system. Among these chronic disorders is irritable bowel syndrome. This disease can be diagnosed by careful evaluation of clinical signs, medical records and simple tests that only require a lot of time and high knowledge. On the other hand, data mining is one of the multidisciplinary fields derived from scientific fields such as statistics, mathematics, computers and artificial intelligence, whose applications are expanding in various fields such as health and treatment.

Applying data mining on medical data brings vital, valuable and effective achievements and can enhance the physician's knowledge to take the necessary measures and also speed up the diagnosis process. Therefore, in this study, the best data mining algorithm for the diagnosis of irritable bowel syndrome was determined using WEKA software. For this purpose, the data mining process using Decision Stump, Random Tree and j48 algorithms on 59 samples collected from the files of people referring to Erfan Hospital in Tehran (to collect information about patients) and questionnaires distributed among employees of a company. Software was used to collect information from healthy individuals. Finally, the total data included 55.93% male sample and 44.06% female sample. The results were compared in terms of speed, accuracy and computational cost. Finally, j48 algorithm with 96.6% accuracy was introduced as the optimal algorithm for IBS detection.

Keywords: Irritable Bowel Syndrome, Data Mining, j48, Random Tree, Decision Stump

1- Introduction

Advances in data collection and storage capabilities in recent decades have led to large data volumes in many sciences. Statistical techniques and traditional management tools are not enough to analyze this data, therefore extracting knowledge from this large amount of data is a big challenge. On the other hand, in medical science, collecting a lot of data about each disease is of special importance that obtaining results from this data, helps a lot to diagnose, analyze, prevent and treat some diseases. But the problem is the large amount of data that confuses and prevents usable and useful results. In fact, large data cannot be used alone, but the knowledge hidden in the data can be used, which can be used with the help of data mining.

Irritable bowel syndrome (IBS), sometimes referred to as Crohn's disease, is one of the diseases for which a lot of data must be studied to diagnose. Irritable bowel syndrome is very common and affects about 10 to 20% of the population [1,2]. After a cold, the disease is the most common cause of medication. These patients make up more than 50% of referrals to internal medicine physicians. Many doctors ask the patient for the results of colonoscopy and serology at the very beginning of the patient's visit to save time in diagnosing the disease, which will definitely cost the patient a lot. The aim of this article is to find the best data mining algorithm to diagnose irritable bowel syndrome using patient data analysis.

Medical data mining in addition to the technical issues that exist in other areas of data mining, it faces some non-technical problems [3]. There are ethical, legal, and social constraints on the collection and distribution of medical data that do not exist on other data, and this limits the conclusions that can be drawn. In other words, medical data have unique characteristics, some of which are: heterogeneity of medical data, ethical, legal and social issues that exist about this type of data [4].

1-1 Irritable Bowel Syndrome (IBS)

Gastrointestinal diseases are one of the most important and common chronic non-communicable diseases. Among these chronic disorders is irritable bowel syndrome. This syndrome is an intestinal dysfunction that is characterized by changes in bowel movements and abdominal pain in the absence of identifiable structural abnormalities that can significantly affect the patient's quality of life [5]. There is no test to definitively diagnose IBS.

1-2 Diagnostic criteria of the disease

Since this disease does not have a specific structural disorder or biochemical sign, its diagnosis is based on the presence of clinical signs.

In order to create more coordination in the reporting of symptoms and increase diagnostic accuracy, clinical criteria for diagnosis were presented. The first criteria were proposed in 1987 by Manning et al. to standardize clinical research protocols. An international working group provided a comprehensive definition in 1990 called the Rome I criteria, which were revised in 1999 under the name Rome II[6].

The American Gastroenterologists Association (AGA) recommends that the diagnosis of IBS should be based on the presence of symptoms consistent with the disorder (Rome criteria) along with the rejection of other disorders that cause similar clinical manifestations.

The rest of paper is organized as follows. In section 2, previous related works have been reviewed. In the third section, the definition of the problem, the method of knowledge discovery as well as the description of the algorithms have been introduced, and in section 4 experimental results is presented. The conclusion is discussed in section 5.

2- Related Works

Gastrointestinal diseases are one of the most important causes of death and the number of patients with it has an increasing trend. Irritable bowel syndrome is also a gastrointestinal disorder with a wide range of symptoms. Despite the research that has been done in this case, but the main cause of this disease is still unknown and is considered as a multi-causal problem. Factors such as family history, maladaptation between the nervous system and the gastrointestinal tract, anxiety, lifestyle and diet are known as factors influencing the onset and exacerbation of symptoms. According to studies, the risk of other gastrointestinal diseases in patients with irritable bowel syndrome increases up to 8 times.

Among the main symptoms of this syndrome are changes in bowel movements and pain in the abdomen. The prevalence of this disease is reported between 10 and 20% in the total population and in Iran about 6% and is one of the main reasons for people to go to medical centers and absence from work, so that about 30% of people seek the symptoms of this disease refer to a hospital. As a chronic disorder, this syndrome affects various aspects of patients' lives, including social relationships and their job performance and imposes many financial costs on society and the patient. The cost of treating this disorder is estimated at about $ 2.8 million in urban areas, which puts a significant financial burden on the country's economy. According to studies, the prevalence of this disorder is higher in women than men, and women make up 80% of people with severe symptoms. In 50% of patients, the symptoms appear before the age of 35. According to published statistics, about 10 to 15% of people in Europe and North America also have this syndrome. Correct diagnosis of this disease is important and today Rome-III criteria are used for diagnosis, but a definite opinion about the patient is made only after ensuring that he does not have other disorders and diseases, as a result of which the person incurs costs for tests [7].

Information collection and management is of great importance in the medical sciences because it plays a key role in the timely diagnosis or prediction of diseases. One of the newest methods in this field is data mining, which means searching for extensive information resources and achieving meaningful communication and patterns that statistical analysis has not been able to detect and can reduce costs and improve quality in the field. Medicine is useful [8]. In an article presented by Penny et al. [9], they used data mining to identify health-related quality of life indicators in patients with IBS. The quality of life of people with IBS symptoms was questioned and then assessed by a valid questionnaire and several data mining models were used to assess the factors affecting the quality of life of people including logistic regression, classification tree and artificial neural networks. The results showed that psychological and

social complications such as marital status and employment status, etc. are directly related to the severity of gastrointestinal symptoms. In an article presented by Jothi et al.[10], articles that have tried to diagnose the disease with the help of data mining have examined the results, methods and algorithms used. The findings of this paper indicate that only a specific method in data mining cannot be suitable for problem solving (disease diagnosis), but to obtain the highest accuracy, a hybrid method can be used to be able to solve problems in the best possible way. In the way suggested by Kharya [11], data mining techniques have been used to diagnose and predict breast cancer. Some effective techniques that can be used to classify breast cancer are discussed. Among the information classification methods and calculation methods, the decision tree with 93.62% accuracy is the best predictor and detector of cancer.
In the article by Mahmoudi et al. [12], the effective factors in cancer incidence have been investigated using data mining. Data mining algorithms have been implemented using MATLAB software and two decision tree algorithms ID3 and Apriori have been used and it has been concluded that ID3 algorithm has a high model power in predicting the probability of disease in the individual. This article has suggested the use of feature selection methods and its combination with the existing methods in the research for future work and promotion of the research. In an article presented by Rafeh at al. [13], they have studied and used data mining techniques to diagnose diabetes. Used. This research was performed on the data of 5706 patients and then after modeling using different classification techniques, the C4.5 decision tree was selected as the best data mining model with 90.02% accuracy. In an article presented by Abdar et al.[14], they compared the performance of data mining algorithms in predicting previous diseases. The aim of this study was to compare different data mining algorithms on predicting heart disease. This work uses data mining techniques to predict the risk of heart disease. After feature analysis, the models were developed and validated by five algorithms including C5.0, neural network, SVM, KNN and logical regression. C5.0 The decision tree was able to build the model with the highest accuracy of 93.02%, KNN, SVM, neural network 88.37%, 86.05% and 80.23%, respectively. The results of the decision tree can be easily interpreted and implemented. In the method proposed by Dalavi et al.[2], the demographic, social and clinical characteristics of patients with irritable bowel syndrome have been investigated. The sampling method was purpose-based and entered into analysis using SPSS software. Descriptive statistics were used to describe the subjects and analysis of variance was used to compare the variables. This article can be of great help in preparing IBS clinical signs in this dissertation.

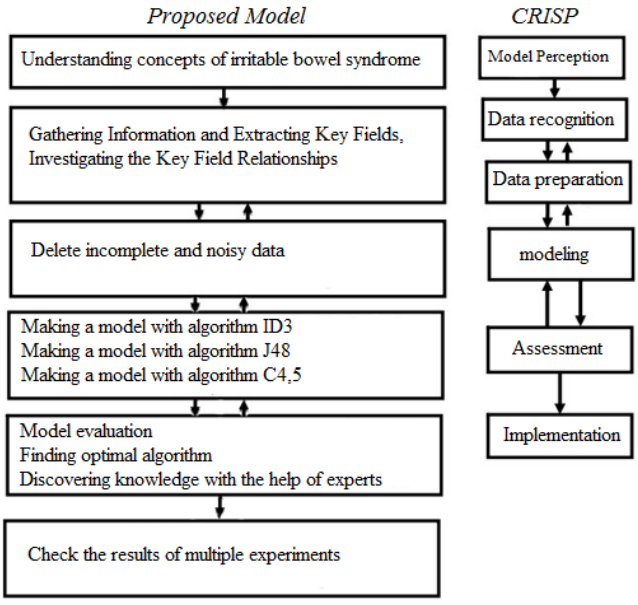

Figure 1. CRISP method steps and proposed model

3- Proposed Method

There are many methods for implementing data mining projects, one of which is the CRISP (Cross Industry Process for Data Mining) methodology. In this research, the proposed model based on CRISP is presented which includes 6 phases: 1) Business understanding, 2) Data understanding, 3) Data preparation, 4) Modeling, 5) Evaluation, 6) Deployment [15].
As shown in Figure 1-3, forward and backward movement between phases is necessary because the input of each phase depends on the phase output of the previous stage. The direction of the arrows in the figure shows this dependence well. The remarkable point in this model is that it is possible to go back to any of the previous steps and apply the necessary corrections in it. For example, in the modeling phase, it may be necessary to return to the preparation phase and make changes to the data set, depending on the circumstances.
In this paper, one data miner, one nurse and two physicians have collaborated to prepare the questionnaires and complete and review their results. To collect the data of this research, a questionnaire was designed under the supervision of a physician, which was finally reviewed and approved by 2 other hospital physicians. This questionnaire includes a summary of the questions asked in ROME-III and a set of questions that the doctor asks to rule out other diseases from the client. Data processing as well as their modeling has been done with the help of WEKA software version 3.9 and the purpose of this data mining is to compare J48, Decision Stump, Random Tree algorithms to find and introduce the optimal algorithm among them.

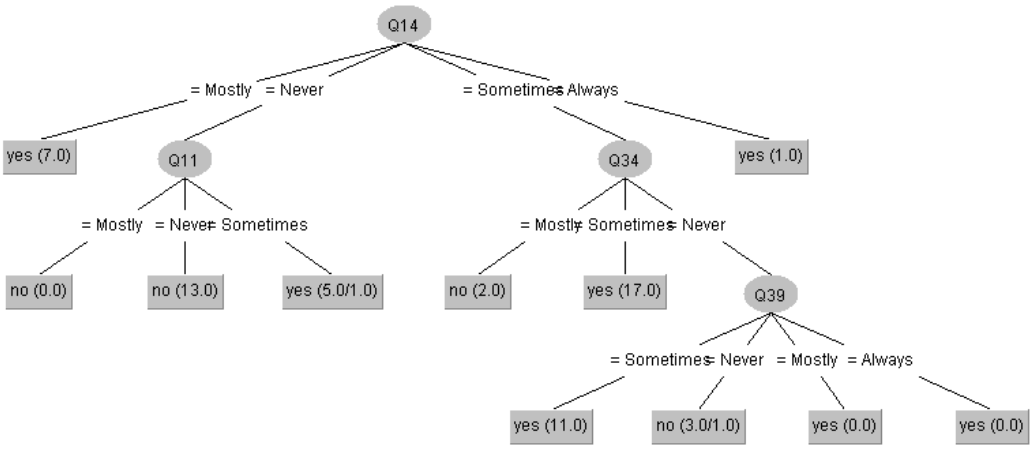

Figure 2. J48 tree

4- Experimental Results

The data set of this study consists of 2 categories, of which 30.5% of the data (18 people) are related to healthy people and 69.49% of the data (41 people) are related to sick people. Data related to healthy individuals were collected by distributing a questionnaire among the employees of a computer company and data related to sick individuals by reviewing the medical records of individuals referring to Erfan Hospital clinic in Tehran with gastrointestinal disorders.
One of the important points at this stage is data quality verification. Issues such as data dimensions, plug-in data, incompatible data, and incomplete data need to be identified and addressed. For this purpose, a number of data were merged, which is essential for high-volume data.
Incomplete data is data that does not have enough information to extract knowledge, and plug-in data is data that is not important in the conclusion and can be identified and deleted by feature selection and feature extraction algorithms. They showed the severity of some of the questions. Due to the lack of positive effect and neutrality of these cases, they were identified as add-on data by a physician who specializes in diagnosing the disease and were removed manually.

Examining the histogram of each profile, we observed that a number of questions have a value of Null, which indicates that the person did not answer the relevant question. Due to the negative impact of these questions on the output, the questions were cleared.

4-1 Modeling using the J48 algorithm

In this part, modeling was performed using the J48 decision tree algorithm, the last question that indicated whether or not a person had the disease was selected as the classification variable. 10 times and each time 1/10 of the data should be used as a test set to evaluate the model. At the end, the average of these 10 executions was displayed as the final output. Or the model can be tested with the same set of educational data, which of course will provide us with the most ideal type of evaluation, because in this case all the data are properly classified in their classes.

The decision tree obtained from this algorithm, as shown in Figure 2, is that first the tree is created using the variable Q14 (question 14 in the questionnaire) and then in the next step, from the variables Q11, Q34 and Q39 (to Questions 34, 11 and 39, respectively) are used to continue the work. Finally, the number of leaves of the tree is 11 and the total number of nodes is 15 and the time elapsed to create the tree is 0.03 seconds. The accuracy of this method was calculated to be 96.6%.

4-2 Modeling using Random Tree algorithm

In this section, modeling was performed using the Random tree decision tree algorithm. The last question that indicated whether or not a person had the disease was selected as the classification variable. The model was tested with the same set of educational data entered.

The decision tree obtained from this algorithm, as shown in Figure 3, is that first the tree is created using the variable Q23 (question 23) and then in the next step, the variables Q22, Q26, Q8, etc. for Continued work is used. Finally, the generated tree contains 29 nodes and the time elapsed to create the model is calculated to be 1.18 seconds. The accuracy of this method was calculated to be 88.13%.

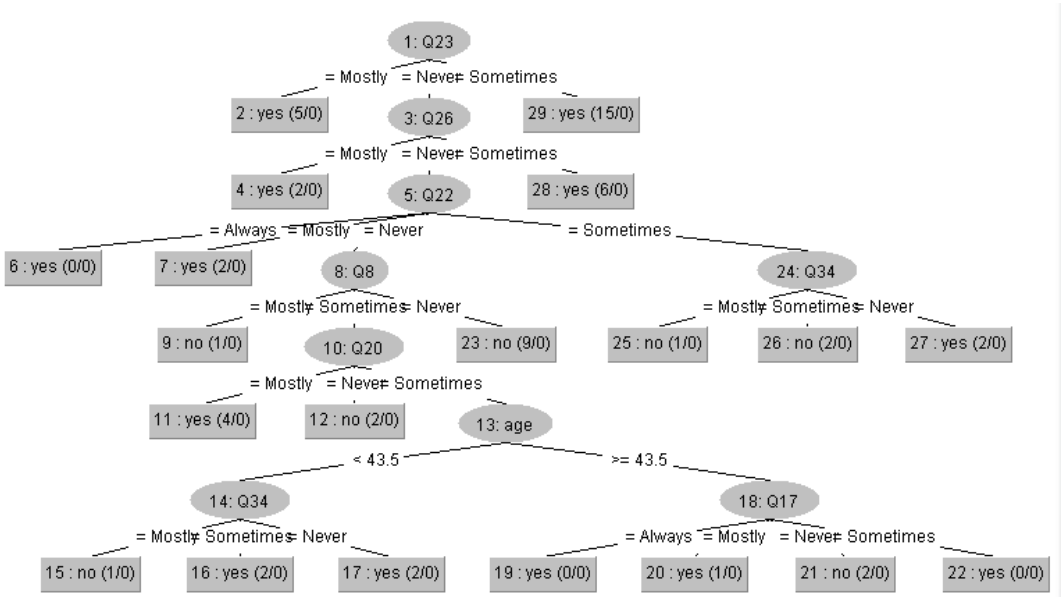

Figure 3. Random Tree algorithm

4-3 Modeling using Decision Stump algorithm

In this section, modeling was performed using the Decision Stump algorithm, again the last question that indicated whether or not a person had the disease was selected as the classification variable. The model was tested with the same set of educational data entered. The accuracy of this algorithm was 91.52%.
Figure 4 shows the accuracy of the classifications obtained from the algorithms applied to the data set described in this research. The J48 algorithm with 96.6% accuracy as the algorithm with the highest accuracy and then the Decision Stump algorithms with 91.52% accuracy and Random Tree algorithm with 88.13% accuracy can be seen in the mentioned diagram.

5- Conclusion

The results of this study indicate that in the data set that was examined, there is knowledge that can be used to achieve early diagnosis of this disease with minimal cost and time. From the results obtained from the classifications and application of algorithms used on the data, J48 algorithm can be introduced as the optimal algorithm in terms of accuracy and speed of diagnosis of this disease with 96.6% accuracy compared to other algorithms.
Also, the questions set in the questionnaire are designed in such a way that it includes questions related to ROME-3 as well as a selection of important and key questions to rule out other gastrointestinal diseases that can be helped by physicians for early diagnosis. Sickness and time savings as well as cost savings to patients for not having to undergo tests and other costly medical procedures.

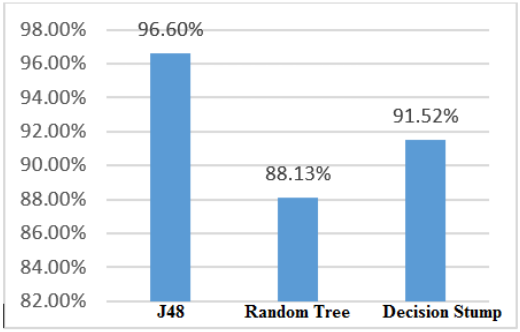

Figure 4. Comparison of the accuracy of classification algorithms

References

1. Wilson S, Roberts L, Roalfe A, Bridge P, Singh S (2004) Prevalence of irritable bowel syndrome: a community survey. Br J Gen Pract 54 (504):495-502
2. Khajedaluee M, Vosooghinia H, Bahari A, Khosravi A, Esmaillzadeh A, Ganji A, Akhavan Rezayat K, Mahmoudi R (2014) Demographic, social and clinical characteristics in patients with irritable bowel syndrome in Mashhad in 2013. medical journal of mashhad university of medical sciences 57 (3):579-586
3. Hasanpour H, Meibodi RG, Navi K (2019) Improving rule-based classification using Harmony Search. PeerJ Computer Science 5:e188
4. Lavrač N (1999) Selected techniques for data mining in medicine. Artificial intelligence in medicine 16 (1):3-23
5. Malekzadeh R, Derakhshan MH, Malekzadeh Z (2009) Gastric cancer in Iran: epidemiology and risk factors.
6. Longstreth GF, Thompson WG, Chey WD, Houghton LA, Mearin F, Spiller RC (2006) Functional bowel disorders. Gastroenterology 130 (5):1480-1491
7. Ghadir MR, Ghanooni AH (2014) Review of pathophysiology and diagnosis of irritable bowel syndrome. Qom University of Medical Sciences Journal 7 (6):62-70
8. Gholamhosseini L, Damroodi M (2015) Evaluation of data mining applications in the health system. Paramedical Sciences and Military Health 10 (1):39-48

9. Penny KI, Smith GD (2012) The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome. Journal of clinical nursing 21 (19pt20):2761-2771
10. Jothi N, Husain W (2015) Data mining in healthcare–a review. Procedia computer science 72:306-313
11. Kharya S (2012) Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint arXiv:12051923
12. Mahmoodi SA, Mirzaie K, Mahmoodi SM (2017) Determining the effective factors in the incidence of gastric cancer by using data mining approach. Journal of Payavard Salamat 11 (3):332-341
13. Rafeh R, Arbabi M (2015) Data mining techniques to diagnose diabetes using blood lipids. scientific journal of ilam university of medical sciences 23 (4):239-247
14. Abdar M, Kalhori SRN, Sutikno T, Subroto IMI, Arji G (2015) Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of Electrical & Computer Engineering (2088-8708) 5 (6)
15. Elmi Sola Y, Pourjavad MA, Hasanpour H, Zojaji H, Analoui M Load balancing effects in DCUR QoS routing algorithm. In: 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009. IEEE, pp 210-213