International Journal for Modern Trends in Science and Technology, 8(08): 149-152, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0808020
Available online at: http://www.ijmtst.com/vol8issue08.html

Prediction and Analysis of Wilson Diseases using Hepatitis Data Sets

Gomathy G, Kalaiselvi P, Jena Catherine Bel D, Zionna Sen G B

Department of Artificial Intelligence and Data Science, Sri Sai Ram Engineering College, Chennai

To Cite this Article
Gomathy G, Kalaiselvi P, Jena Catherine Bel D and Zionna Sen G B. Prediction and Analysis of Wilson Diseases using Hepatitis Data Sets. International Journal for Modern Trends in Science and Technology 2022, 8(08), pp. 149-152. https://doi.org/10.46501/IJMTST0808020

Article Info
Received: 11 July 2022; Accepted: 06 August 2022; Published: 12 August 2022.

ABSTRACT

Wilson Diseases is a genetic disease prevents the body from removing extra copper. It is an autosomal recessive disease. Wilson Diseases is a rare autosomal recessive inherited disorder. The objective of this research is to predict or detect / find Wilson syndrome genetic disease from the gene and the goal is to reduce copper in the body or remove the excess copper content in the body. The chance of getting this disease is about 1 to 40000 people worldwide. In this paper we apply hepatitis dataset which mainly concentrated on predicting Wilson syndrome gene disease. For prediction some of the important data mining techniques are used like classification, clustering, association rules, mining frequent pattern, correlations, prediction etc. Disease prediction plays an important role in data mining. In this paper we apply different classification algorithms are used to find the accuracy.
Keyword: Wilson, Classification, Prediction, Hepatitis, accuracy, Data Mining.

1. INTRODUCTION

Data Mining is a process of Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases. Data Mining technique support automatic Exploration of data and attempts to source out patterns and trends in the data. Data Mining refers to extracting or mining knowledge from large amounts of data. Sifting through very large amounts of data for useful information. Data mining tasks are Prediction and Description. Prediction use some variables to predict unknown or future values of other variables. Common tasks used in Prediction Classification, Regression, Deviation Detection Description is used to find the human interpretation patterns that describes the data. The tasks are Clustering, Association Rule, Sequential pattern Discovery.

2. ESTIMATING ACCURACY

Estimating Accuracy is the important role in the data mining. It is used for to calculate how the algorithm is compared with the given dataset and it produce the relevant result. In this concept the attribute selection is based on the Best first search. It starts from in visited vertex. Then all unvisited vertices vi adjacent are visited and then all unvisited vertices wj adjacent are visited. The traversal terminates when there are no more nodes to visit. It uses a queue data structure to keep track of order of nodes whose adjacent nodes are to be visited.

3. IMPLEMENTATION

Choose any node in a graph, designate it as the search node & mark it has visited. Using the adjacency matrix of the graph, find all the unvisited adjacent nodes to the search node and enqueue them into the queue Q. Then the node is dequeued from the queue. Mark that node as visited and designate it as the new search node. Repeat step2 and 3 using the new search node. This process continues until the Queue Q which keeps track of the adjacent nodes is empty.

4. DATASET DESCRIPTION

The hepatitis dataset collected from UCI repository. About the hepatitis database and BILIRUBIN problem I would like to say the following: BILIRUBIN is continuous attribute (= the number of its "values" in the ASDOHEPA.DAT file is negative!!!); "values" are quoted because when speaking about the continuous attribute there is no such thing as all possible values. However, they represent so called "boundary" values; according to these "boundary" values the attribute can be discredited. At the same time, because of the continuous attribute, one can perform some other test since the continuous information is preserved.

5. ATTRIBUTE DESCRIPTION

Attribute contains 155 instance and 20 number of attributes are implemented. Including the class attribute. A class have two function one is DIE another one is LIVE. The attributes are, CLASS, AGE, SEX, STEROID, ANTIVIRALS, FATIGUE, MALAISE, ANOREXIA, LIVER_BIG, VER_FIRM,SPLEEN_PALPABLE,SPIDERS,ASCITES,VARICES,BILIRUBIN,ALK_PHOSPHATE, SGOT, ALBUMIN, PROTIME, HISTOLOGY.

The dataset used in this concepts Wilson syndrome dataset. It is available in UCI Repository. It is in the form of .arff file format. Goal is to reduce copper in the body, remove the excess copper content in the body. The chance of getting this disease is about 1 in 40000 people worldwide. Appears between ages 6 to 40 but can starts as early as 2 and as late as 72. Feature of this condition include a combination of liver disease and neurological and psychiatric problems. It occurs in all varieties of people, but most common in European, Sicilians, Italians etc. Symptoms begin to show by age 4. The Symptoms are Vomiting blood, Enlargement of the abdomen, Confusion and delirium, Jaundice. Diagnosing eye Examination, physical text, Lab test etc.

6. EXPERIMENT RESULT

Attribute Selection on all input data. The Search Method is Best first. The set starts with the no of attributes. Start set: no attributes Search direction: forward Stale search after 5 node expansions. Total number of subsets evaluated: 189. Merit of best subset found: 0.323. Attribute Subset Evaluator (supervised, Class (nominal)20 Class) CFS Subset Evaluator. Including locally predictive attributes, Selected attributes are 1,2,6,11,12,13,14,17,18,19: Total 10 , AGE, SEX, MALAISE, SPIDERS ASCITES, VARICES, BILIRUBIN, ALBUMIN, PROTIME, HISTOLOGY.

6.1 clustering analysis

Clustering model (full training set). Number of cluster selected by cross validation: 5 Number of iterations performed: 11. Time taken to build model (full training data) : 1.94 seconds Number of clusters selected by cross validation: 5 , Number of iterations performed: 13. Time taken to build model (percentage split) : 0.96 seconds

6.2 Clustered Instances

0	15 (28%)
1	6 (11%)
2	15 (28%)
3	8 (15%)
4	9 (17%)

6.3 Classification analysis

Classifier model (full training set) ZeroR predicts class value: LIVE. Time taken to build model: 0 seconds Evaluation on training set, Time taken to test model on training data: 0.01 seconds Algorithm: CfsSubsetEval attribute Selection BestFirst

Original Attribute:

L={CLASS, AGE, SEX, STEROID, ANTIVIRALS,FATIGUE,MALAISE,ANOREXIA,LIVER_BIG,VER_FIRM,SPLEEN_PALPABLE,SPIDERS,ASCITES,VARICES,BILIRUBIN,ALK_PHOSPHATE,SGOT,ALBUMIN,PROTIME, HISTOLOGY.}

List of usage of Attributes 90%---100%

Attribute Name	Cross fold validation
AGE	60%
SEX	100%
STEROID	0%
ANTIVIRALS	10%
FATIGUE	50%
MALAISE	60%
ANOREXIA	0%
LIVER_BIG	0%
LIVER_FIRM	0%
SPLEEN_PALPABLE	20%
SPIDERS	100%
ASCITES	100%
VARICES	100%
BILIRUBIN	100%
ALK_PHOSPHATE	0%
SGOT	0%
ALBUMIN	100%
PROTIME	100%
HISTOLOGY	80%

Table : 1 Classification Results

weka.attributeSelection.CfsSubsetEval

This table is derived from implementing CfsSubsetEval attribute Selection BestFirst in weka tool. It shown how its correctly classified a different attributes and display the cross fold validation results. Some of the Attributes are correctly classified 100 % percentage and few is classified as 80 %, 60 % etc.

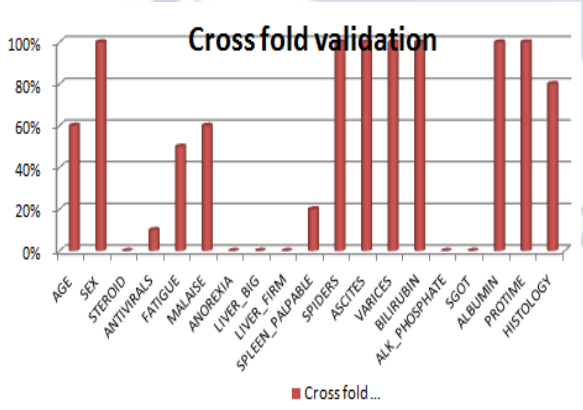

Fig: 1 Shows the graphical representation of difference in Accuracy

Attribute Selection on all input data

Search Method:

Best first algorithm is used, start set no attributes, Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 189, Merit of best subset found: 0.323 Attribute Subset Evaluator (supervised, Class (nominal): 20 Class): CFS Subset Evaluator Including locally predictive attributes

Selected attributes are (1,2,6,11,12,13,14,17,18,19 : 10)

Table : 2 Classification Results

Sno.	Attribute Name	Percentage
1	SEX	100
2	SPIDERS	100
3	ASCITES	100
4	VARICES	100
5	BILIRUBIN	100
6	ALBUMIN	100
7	PROTIME	100

Table: 3 Attribute and Percentage results

Correctly classified Instance	**123**	**93.3548**
Incorrectly Classified Instance	**32**	**7.6452**
Kappa Statistic	**0**	**-**
Mean Absolute Error	**0.3299**	**-**
Root mean Squared error	**0.4048**	**-**
Relative absolute error	**-**	**100%**
Root relative squared error	**-**	**100%**
Total Number of Instance		**155**

7. CONCLUSION

Genetic problem; Wilson disease is one of an inherited disorder possibly to a life threatening level.However proper treatment taken when diagnosed early may increase the life span. There are variousmajor data mining techniques developed and used which includes association, clustering,classification, prediction, sequential patterns and regression. . Hepatitis datasets usingcross validation technique the performance of

attributes are estimated for accuracy while compiling.Best First and CFS subset event is used for attribute Selection.In this concept classification is used toobtain relevant and accurate information about data from different classes and achieved 93 % accuracy.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] S. Vijiarani , S. Sudha, "Disease Prediction in Data Mining Technique", International Journal of Computer Applications and Information Technology, Vol. II, Issue I , January 2013(ISSN:2278-7720).

[2] Ferenci P, Caca K, Loundianos G, Mieli - Vergani G, Tanner S, Sternlieb I, Schilsky M, Cox D, Berr F. Diagnosis and phenotype classification of wilson disease . Liver International 2003: 23: 139-142, ISSN 1478-3223.

[3] S. Tharageshwari, D. Sasikala "An Algorithm to Detect Kayser-Fleischer Ring in Human Eye for Diagnosing Wilson Disease" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 5, May 2014.

[4] A Review and Current Perspective on Wilson Disease Mallikarjun Patil, Keyur A. Sheth, Adarsh C. Krishnamurthy, and Harshad Devarbhavi∗v.3(4); 2013 Dec PMC3940372

[5] Wilson's disease: A review of what we have learned.Rodriguez-Castro KI, Hevia-Urrutia FJ, Sturniolo GC. World J Hepatol. 2015 Dec 18;7(29):2859-70. doi: 10.4254/wjh.v7.i29.2859.

[6] The onsetof psychiatric disorders and Wilson's disease15-17, rue du Clos-Bénard, EPS de Ville-Evrard, secteur 93G06, 93300 Aubervilliers, France.

[7] "Diagnosis and Treatment of Wilson Disease: An Update" , Eve A. Roberts and Michael L. Schilsky. AASLD PRACTICE GUIDELINES.

[8] Mak C, Lam C: Diagnosis of Wilson's disease: A comprehensive review. Crit Rev Clin Lab Sci 2008;45:263–290.

[9] Noble J: A case study: identifying a new case of Wilson's disease. J Am Acad Nurse Pract 2005;17:512–517.

[10] Brewer G, Dick R, Johnson V, Brunberg J, Kluin K, Fink J: Treatment of Wilson's disease with zinc: XV long-term follow-up studies. J Lab Clin Med 1998;132:264–278.

[11] Sakaida I, Kawaguchi K, Kimura T, Tamura F, Okita K: D-penicillamine improved laparoscopic and histological findings of the liver in a patient with Wilson's disease: 3-year follow-up after diagnosis of Coombs-negative hemolytic anemia of Wilson's disease. J Gastroenterol 2005;40:646–651.

[12] Hlubocká Z, Mareček Z, Linhart A, et al: Cardiac involvement in Wilson's disease. J Inherit Metab Dis 2002;25:269– 277.

[13] Ricciardi M, Sirimarco G, Vicenzini E, et al: Transcranial sonographic findings in Wilson's disease. J Ultrasound Med 2010;29:1143–1145.