Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230736

Diagnosis of hepatitis C virus based on ID3 algorithm

Yue Che

Dalian Maritime University, Dalian, China, 116026

CheYue7727@outlook.com

Abstract. Infection with the hepatitis C virus (HCV) is what causes hepatitis C. China has one of the highest hepatitis C infection rates in the world (13%-15%), and this illness affects more than 170 million people worldwide. As a result, there is an enormous need for the diagnosis of this disease. In recent years, researchers have made significant progress in this field with different machine learning algorithms and have had the ability to make a relatively precise diagnosis. However, within the machine learning algorithms that the researchers used, decision trees, particularly the ID3(Iterative Dichotomiser 3) algorithm, have been disregarded. This paper uses this algorithm in the diagnosis of hepatitis C, aiming to test the accuracy and explore the possibility of introducing this approach into application in medicine. Empirically, the novel method of applying the ID3 algorithm in the diagnosis of hepatitis C can produce relatively accurate results compared with those traditional approaches, demonstrating that this method can be used in practice and is a powerful approach for diagnosing people who have the hepatitis C virus.

Keywords: hepatitis C, machine learning, ID3 algorithm, decision tree.

1. Introduction

Parentally transmitted non-A/non-B hepatitis is primarily caused by the Hepatitis C virus [1]. With the development of science and technology, hepatitis C is able to be diagnosed and then be cured nowadays. However, the hepatitis C virus is not eliminated, and there are numerous people who are infected but not diagnosed around the world. It will result in liver failure, cirrhosis, and liver cancer if it is not treated in a timely manner. Therefore, the early detection for the hepatitis C virus is a huge problem that people have to solve, and successfully finding a viable method to make the diagnoses is a great contribution to medical science.

Machine learning algorithms, such as random forests, k-Nearest Neighbor(KNN) and CART(classification and regression tree), have already been applied as powerful means to tackle this issue, which allows the computers to analyse and process the data from a database and produce diagnoses based on the results. As a representative machine learning algorithm, the ID3 algorithm should have been considered as an approach to making diagnoses for the hepatitis C virus. However, earlier researchers overlooked its potential, because the ID3 algorithm is designed for processing discrete data, but most data stored in medical science databases is continuous one. As a result, few researchers have focused on this field, which has caused the research gap in this area.

One of the primary causes of liver cirrhosis and hepatocellular cancer globally is chronic infection with the hepatitis C virus [3]. Since there is such an urgency to make early diagnoses for hepatitis C and

DOI: 10.54254/2755-2721/6/20230736

the ID3 algorithm has the potential to finish this challenging work, this paper tries to assess the ability of the ID3 algorithm in making diagnoses by calculating the accuracy of classification and making comparisons with alternative algorithms for purpose of determining the performance of the ID3 algorithm. In addition, the goal of this study is to serve as an example for other scholars who are keen to investigate the same topic.

2. Literature review

In recent years, tools for prediction, classification, and diagnosis have been developed using machine learning approaches including classification trees and artificial neural networks (ANN) [4]. Firstly, it is helpful to review the progress that former researchers have made, as they can be used as criteria to evaluate the accuracy that is produced by the ID3 algorithm in further study, and it is also useful in making comparisons so that the availability of using the ID3 algorithm to make diagnoses can be seen clearly.

In this section, six classic machine learning algorithms are mentioned, as shown in Fig. 1, including Logistic Regression (LR) , Linear Discriminant Analysis (LDA) , k-Nearest Neighbor (KNN) , Classification and regression tree (CART) , Naive Bayes model (NB) and Support Vector Machine (SVM) . These approaches are widely used in making classifications and diagnoses of diseases. Accuracy will be the benchmark for the assessments to be made in order to gauge how well each algorithm performs.

Table 1 shows the accuracy of six machine learning algorithms with arithmetic means and standard deviations. The data used for testing algorithms is from Ainshams University, collected in El Demerdash Hospital. Before testing these algorithms, in order to make the original data equally distributed, the data is preprocessed and normalized. It is clear from the table that KNN and SVM have better performances among the six approaches. The arithmetic means for them are 0.825735 and 0.836397, respectively.

Table 1. The arithmetic mean and the standard deviation of accuracy for each machine learning algorithm [5].

Algorithm	Arithmetic Mean	Standard Deviation
LR	0.734191	0.095885
LDA	0.746324	0.117854
KNN	0.825735	0.054511
CART	0.710662	0.089362
NB	0.648897	0.141868
SVM	0.836397	0.088697

Distributions for classification results need to be taken into account as well in order to determine the effectiveness of each strategy. As shown in figure 1, a boxplot graph is used to demonstrate distributions for the results.

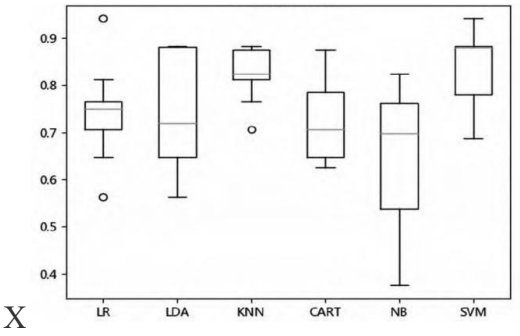

Figure 1. Outcomes of each machine learning algorithm's classification distributions [5].

DOI: 10.54254/2755-2721/6/20230736

As shown in Figure 1, KNN and SVM have the best performances among these machine learning algorithms. KNN can be considered to outperform SVM, due to having the most compact structure, standing at almost 90% accuracy at best and about 76% accuracy at least. Corresponding to Table 1, SVM also has a relatively compact structure, suggesting that it performs better compared with the remaining four approaches.

According to the table and figure above, the criteria for assessing the performance of the ID3 algorithm in making diagnoses for hepatitis C virus are clear. Since k-Nearest Neighbor (KNN) algorithm has the best performance among six traditional approaches, as long as ID3 algorithm is able to produce a better or approximately the same result, then it is sufficient to believe that using ID3 algorithm to make classification for hepatitis C virus is viable.

3. Method

The ID3 algorithm is one of the commonly used decision tree algorithms, which is widely applied in computer science areas such as machine learning, knowledge discovery, and data mining. It is based on the principle of Occam's razor, meaning that 'Entities should not be multiplied unnecessarily'. As a classification prediction algorithm, J. Ross Quinlan initially put forth the ID3 algorithm in 1975.

The ID3 algorithm, a classification system, employs a greedy strategy by choosing the best attribute from a group of attributes that produces the most Information Gain (IG) or the lowest Entropy (H) , because the greater the Information Gain, the stronger the ability to distinguish the sample, and then produce a more precise classification.

Information Gain illustrate the amount of uncertainty in set S that will be cut after using a certain attribute to split the original sample, which can be calculated for each remaining attribute in the ID3 algorithm. Entropy is the characteristic regarding how much uncertainty in a set S. According to the definitions, it is clear that the greater the Information Gain, the better, while the Entropy is expected to be smaller.

The ID3 algorithm's fundamental step is to assess each attribute's potential Information Gain and select the one with the highest Information Gain for splitting. The ID3 decision tree outputs binary classification decisions, with "0" denoting normal and "1" denoting anomaly class assignments to test instances, for the purpose of finding anomalies [6].

The steps in the ID3 algorithm are shown in Figure 2. The process will be reiterated until a complete decision tree is created. In the step of 'judgment', the algorithm needs to assess the tree that has been created. If the tree is already a complete one, then the process is finished, otherwise the process needs to be started again.

The specific steps are as follow: 1. Figure out the sample's entropy. 2. For every attribute or feature: 1). Determine the categorical values' entropy; 2). Figure out the feature's information gain. 3. Determine the feature with the greatest information gain. 4. Reiterate the steps till we have the tree we want.

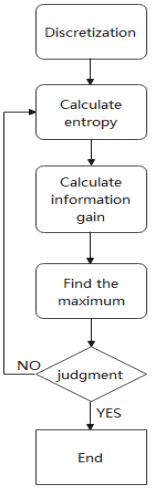

Figure 2. The process diagram for ID3 algorithm.

DOI: 10.54254/2755-2721/6/20230736

Compared with alternative machine learning algorithms, the ID3 algorithm has its special merits. The structure of the ID3 algorithm is simpler, and its flexibility is also remarkable, implying its potential to be used in medical science. However, the ID3 algorithm is not able to process continuous data, which is also the biggest barrier to applying this approach to diagnosing hepatitis C virus.

Overall, the ID3 algorithm is a very useful machine learning algorithm with clear basic theory and strong learning ability, which is suitable for dealing with large-scale machine learning problems and is highly possible to be used in making diagnoses for the hepatitis C virus. It is noticeable that before running the program on the computer, it is necessary to preprocess the data, transforming it from continuous data to discrete data.

4. Data preprocessing and results

4.1. The source of the database

The database utilized for this project is from the Machine Learning Repository of Center for Machine Learning and Intelligent Systems at the University of California Irvine (UCI). This database covers 615 samples in total and the statistical values of 13 different attributes for each of them (As shown in Table 2), including Age, Sex, Albumin (ALB), Alkaline Phosphatase (ALP), Alanine Amionotransferase (ALT), Aspartate Amino Transferase (AST), Bilirubin (BIL), Cholinesterase (CHE), Total Cholesterol (CHOL), Serum Creatinine (CREA), γ - Glutamyltranspeptidase (GGT) and Protein (PROT). In the database, there are a certain number of missing values among these samples, and a label is recorded for each of them in order to make a classification, including 540 blood donors(including suspect blood donors in order to simplify the algorithm), 24 Hepatitis, 21 Fibrosis and 30 Cirrhosis, and the last three of them are diseases induced by affection of the hepatitis C virus. For distinguishing different labels, different numbers are used to describe the samples. To be more precisely, 0 stands for blood donor, 1 stands for Hepatitis, 2 stands for Fibrosis and 3 stands for Cirrhosis.

Table 2. Characteristics of the hepatitis C virus [7].

Feature	Description
Age	age
Sex	sex
ALB	albumin
ALP	alkaline phosphatase
ALT	Alanine amionotransferase
AST	Aspartate amino transferase
BIL	bilirubin
CHE	cholinesterase
CHOL	Total cholesterol
CREA	Serum creatinine
GGT	γ - glutamyltranspeptidase
PROT	protein

4.2. Preprocess the data

As mentioned in section 3, before using the ID3 algorithm to make diagnoses for hepatitis C virus, the first problem that needs to be solved is discretization. Since the original data from the database is continuous one, a criterion is needed as a reference to preprocess the data. The criterion used in this essay is shown in Table 3.

DOI: 10.54254/2755-2721/6/20230736

Table 3. Description of the criterion for each attribute for classification.

	Too Low	Excessive	Normal
Age	<30 (young)	>50 (old)	30~50
ALB	<40	>55	40~55
ALP	<50	>120	50~120
ALT	<5	>35	5~35
AST	NA	>40	<=40
BIL	NA	>17.1	<=17.1
CHE	<6.7	NA	>=6.7
CHOL	NA	>5.7	<=5.7
CREA	<44	>106	44~106
GGT	<11	>50	11~50
PROT	NA	>80	<=80

With the criteria in Table 3, the discretization can be finished, and different numbers have also been used to represent different statuses, such as 0 representing 'Too Low', 1 representing 'Normal' and 2 representing 'Excessive'.

In order to fully evaluate the model, cross-validation was used in this work. The cross-validation mechanism extends the training procedure by one more step. [8]. It divides the original sample set into six parts. Five of them represent training data that are used to train the model, enabling it to generalize more effectively in various situations. The model can then be evaluated using the test data. The accuracy of the model can be estimated and recorded using the test data, and eventually, with comparisons to other conventional methods, it will be obvious how the ID3 algorithm acts while making diagnoses for the hepatitis C virus.

As mentioned, a certain number of samples in which some attribute values are missing. Obviously, it will cause enormous waste if these samples are discarded directly. In addition, among these samples, attribute values are missing partially, but the remaining values are still usable. However, considering the fact that these incomplete samples cannot be taken out to calculate entropy and information gain, when finishing this part of work, incomplete samples are extracted first, and the number of them is recorded at the same time, and then the complete samples are used for calculating entropy and information gain normally. Eventually, the information gain for each attribute is multiplied by a distinct weight, which is the ratio of complete samples (regarding to this certain attribute) to all samples.

4.3. The final result

Because the training data is selected randomly from the original sample set according to the cross-validation, the decision trees built in the ID3 algorithm are distinct for each turn round. Figure 3 shows the structure of one decision tree constructed in a random round.

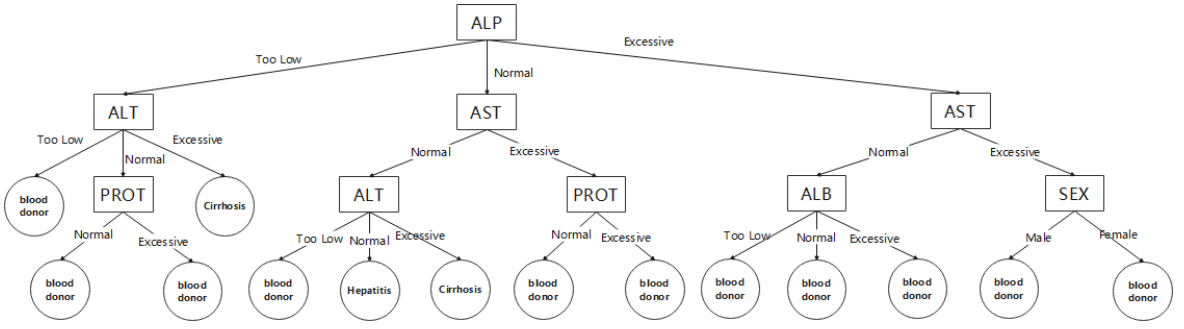

Figure 3. The structure of a random decision tree built by ID3 algorithm.

DOI: 10.54254/2755-2721/6/20230736

Using top-down divide-and-conquer strategy, decision trees perform supervised categorization [9]. In Figure 3, the rectangular modules represent the decision nodes and the circle modules represent the leaf nodes. It is clear from the picture that the decision tree is constructed by four layers, with the attribute 'ALP' acting as the root node, and then is divided into three brunches according to the criteria mentioned in Table 3. Furthermore, the four labels used as the result for classification can be seen in the leaf nodes.

The easiest aspect to assess when judging the model is accuracy. It is possible to record the labels from the test data and the classification outcomes generated by the model, and then use those records to determine the accuracy. The form of the calculation of the accuracy can be shown as:

$$\text{Accuracy} = \frac{\text{The Number of Correct Classification}}{\text{The Number of Samples in Test Data}}$$

Following the equation above, the accuracy of the model can be calculated. Table 4 demonstrates the arithmetic mean and the standard deviation.

Table 4. Arithmetic mean and standard deviation of the accuracy for ID3 algorithm model.

Algorithm	Arithmetic Mean	Standard Deviation	Maximum	Minimum
ID3	0.877348	0.028076	0.932039	0.823529

These four values in Table 4 (Arithmetic Mean, Standard Deviation, Maximum and Minimum) are calculated based on the classification accuracy of 50 distinct rounds, with the arithmetic mean at 0.877348 and the standard deviation at 0.028076.

5. Discussion

After calculating the accuracy of the ID3 algorithm model and constructing the structure of decision tree, this sector will discuss the feasibility of applying ID3 algorithm in diagnoses of hepatitis C virus and the disadvantages that needs to be improved in the future.

Even though the database used in this paper and the database used in Table 1 are not the same, the results in Table 1 can still be used as a helpful reference to evaluate how well the ID3 algorithm model performs.

Firstly, it is noticeable from Table 1 and Table 4 that the ID3 algorithm has a relatively high arithmetic mean and a small standard deviation compared with other machine learning algorithms. As mentioned above, despite the databases being different, they still provide evidence that illustrates the ID3 algorithm has the ability to make diagnoses for hepatitis C virus.

Secondly, it is also clear from Table 4 and Figure 1 that the discrepancy between the maximum and the minimum accuracy in the ID3 algorithm is rather small, which tells us that the result produced by the ID3 algorithm has a compact structure, and the ID3 algorithm has the ability to produce a relatively stable result. That supports the hypothesis of using the ID3 algorithm for detecting and classifying the hepatitis C virus again.

However, according Figure 3, the structure of the final decision tree is only consisted of four layers, with the root node on the top and the leaf nodes on the bottom. Because another drawback of the ID3 algorithm is lacking of strategy to cut off brunches that may lead to overfitting of the model. Always present is the "overfitting" problem, which occurs when training data that a simple model might have fit is instead fitted by a model that is unnecessarily complex [10]. The original intention of building such a structure is to avoid this problem. Nevertheless, the accuracy of classification may be higher if another layer can be added to the decision tree and remain the compact structure of the results at the same time.

6. Conclusion

In the area of identifying and classifying the hepatitis C virus, numerous approaches have been used, and this essay presents a brand-new approach. The ID3 algorithm has proved its potential in making diagnoses for the hepatitis C virus, which has relatively high accuracy and a compact structure of its

DOI: 10.54254/2755-2721/6/20230736

results. The model used in this work has overcome two major disadvantages of the ID3 algorithm, having solved the problem of discretization and having made full use of the incomplete samples in the database. However, the drawbacks of the model cannot be ignored. The decision tree makes a compromise by having only four layers for avoiding the overfitting problem, which lowers the classification accuracy. At the same time, in order to make more persuasive comparisons with other approaches, the accuracy of other machine learning algorithms based on the same database should also be calculated. In this paper, the ID3 algorithm model is proved to have the potential to be applied in diagnoses of hepatitis C virus, but cannot be proved to be the best compared with other approaches. In the future, these disadvantages are likely to be tackled by researchers who are also interested in this area. It would be great progress to improve the structure of the decision tree, and test other machine learning algorithms under the same situation. Overall, with the feasibility of applying the ID3 algorithm in the field of detecting and classifying hepatitis C virus, it is possible to apply this approach widely in medical science in the future.

References

[1] Yang ZR. A probabilistic peptide machine for predicting hepatitis c virus protease cleavage sites, IEEE Transactions on Information Technology in Biomedicine, 2007, 11(5):593-595.

[2] Cai JX et al., Fibrosis and inflammatory activity analysis of chronic hepatitis c based on extreme learning machine, 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 19-21 October 2018.

[3] Chiu HJ et al., Hepatitis c virus detection model by using Random Forest, Logistic-Regression and ABC algorithm, 2022, 10:91045-91058.

[4] Hashem S et al., Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis c patients, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 15(3):861-868.

[5] Li Fang, Liu Huaijin, Tian Qing. A comparative study of hepatitis C liver fibrosis prediction algorithms based on machine learning, Journal of Zhaoqing College, 2022, 43(02): 33-42.

[6] Gaddam SR et al., K-means+ID3: a novel method for supervised anomaly detection by cascading k-means clustering and ID3 decision tree learning methods, IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3):345-354.

[7] Ma L et al., Prediction of disease progression of chronic hepatitis c based on XGBoost algorithm, 2020 International Conference on Robots & Intelligent System (ICRIS) , 07 - 08 November 2020.

[8] Domingo JD et al., Cross validation voting for improving CNN classification in grocery products, IEEE Access, 2022, 10:20913-20925.

[9] Li YH et al., Classifiability-based omnivariate decision trees, IEEE Transactions on Neural Networks, 2005, 16(6):1547-1560.

[10] Cong Y et al., Online similarity learning for big data with overfitting, IEEE Transactions on Big Data, 2017, 4(1):78- 89.