ISSN 2616-5775 Vol. 4, Issue 1: 15-19, DOI: 10.25236/AJCIS.2021.040103

Research on Identifying Liver Diseases Based on Mathematical Models

Yishan Lin

Holderness School, 33 Chapel Lane, Plymouth, NH 03264
ylin22@holderness.org

ABSTRACT. *Along with the economic development, people's material living standards have gradually improved and the pace of life has become faster. The pressure of work and study has brought unhealthy lifestyle to many people. Habits due to stress such as alcoholism, overeating, and staying up late have caused more people to contract liver diseases. This article studies how to use mathematical modeling to predict whether a person has liver disease through various data. Our data comes from the data of patients in the North East of Andhra Pradesh, India. The 583 recorded patients include men and women aged four to ninety years old and above, but most of them were middle-aged and elderly men. We used the calculation method of the support vector machine in mathematical modeling, let the computer try the linear and RBF kernel Support Vector Machine, and calculated the prediction boundary with relatively high accuracy. An accurate dividing line can help doctors judge whether a patient has liver disease. By entering the patient's information and comparing it with the dividing line, it can be seen which side of the line he or she falls on. The predicted result shows that the accuracy is about 72.4%. The use of a mathematical model reduces the workload of doctors and also provides convenience for patients to judge their situations.*

KEYWORDS: *Liver diseases, Indian patient, Support Vector Machine, Accuracy*

1. Introduction

The liver is an essential organ for maintaining the normal metabolism of the human body. Liver disease is the general term for all diseases that occur in the liver. The causes of liver disease include viral, bacterial, and parasitic infections, improper diet, drinking, stones, and other diseases that affect the liver. There are many types of liver diseases, and the symptoms and clinical manifestations of different diseases are also different. Laboratory tests can find out whether the liver function is abnormal. Ultrasound examination, CT examination, and magnetic resonance imaging examination can know exactly the shape, boundary, surface smoothness, nature of intrahepatic cysts, and nodules of the liver. Histopathology and cytopathology examination can make a qualitative diagnosis of the punctured material of liver puncture and ascites puncture. Doctors can make a clear diagnosis based on the comprehensive judgment of medical history, clinical manifestations, and auxiliary examinations. For some viral hepatitis, the diagnosis should be made through virus testing. China is a country with a high incidence of chronic hepatitis B, and the annual incidence of tuberculosis is about 1.3 million, ranking second in the world. China's hepatitis B vaccine is used to prevent hepatitis B, and the main target is infants and young children. The World Health Organization (WHO) has issued a series of regularly updated position papers on vaccines and combination vaccines to prevent diseases with global public health impact.

This article mainly introduces the establishment of a mathematical model of liver disease, and the use of this model to analyze the patient's data and determine whether he is infected with liver disease. This machine learning can be applied to the medical workflow. During the diagnosis and treatment of liver diseases, it can help identify liver diseases, assist doctors in diagnosis and reduce the burden on doctors, and improve accuracy and diagnosis efficiency. In some areas where liver disease is prevalent but lacks medical experts, the model used to predict liver disease can help reduce the workload of doctors. Patients may not need to go to the hospital to perform a simple disease analysis according to their conditions. More time can be spent on patients who need treatment.

Several pieces of research relating to the field of liver decease and mathematical models are done. A group of researchers, including Yuanzhong Li, S. Hara, and K. Shimura, used a model consists of a

ISSN 2616-5775 Vol. 4, Issue 1: 15-19, DOI: 10.25236/AJCIS.2021.040103

training process and locating process with AdaBoost histogram classifier to learn the diverse intensity distributions of liver tumor regions. Their model successfully located the boundaries of the tumors[1]. A study done by Stefan Mihai Petrea and others used multiple linear regression (MLR) models with a stepwise method to estimate the concentration of heavy metals, which are hazardous substances, in liver tissues. Significant MLR models were recorded effectively for K, Cu, Zn, and Na in turbot liver tissue [2]. Fu et al. proposed a novel approach to detect Fatty liver disease (FLD) and Heterogeneous liver. A multi-class linear support vector machine was used for classification. The proposed system had an accuracy of approximately 95% of the classification of liver disorders [3].To identify liver fibrosis, Guitao Cao, Pengfei Shi, and Bing Hu designed a novel method to extract liver features based on ultrasound images. Fisher linear classifier and support vector machine is employed to test groups of liver fibrosis images and healthy liver images [4].

To accurately predict whether a patient is ill, we will first collect a series of data about the patient and classify them into the two categories of whether they are ill. Support vector opportunities are then used to find the best dividing line between these two categories, to predict a person's situation by comparing the data of other patients with the location of this dividing line. We will use Python as the programming language, input the obtained data, let the machine learn and use the kernel support vector machine to calculate the best classification boundary, and then calculate the accuracy of this boundary.

2. Models

Support Vector Machine (SVM) is one of the most widely used machine learning methods [5-6]. In the 1990s, support vector machines became the most popular machine learning method at the time due to their extremely high prediction accuracy and the ability to solve nonlinear classification problems. Support vector machines can be divided into linear support vector machines and kernel support vector machines. The former is aimed at linear classification problems, and the latter belongs to non-linear classifiers. The core problem that all linear classifier needs to solve is how to find the best classification boundary to distinguish the two types of samples. Therefore, we have to choose a hyperplane that maximizes the distance to the nearest data point on each side. If such a hyperplane exists, it is called the maximum separation hyperplane. In the N-dimensional space, the classification boundary is an N-1 dimensional space. The advantage of support vector machine is that it can solve high-dimensional problems, can solve machine learning problems under small samples, can handle the interaction of nonlinear features, has no local minimum problem, does not need to rely on the entire data, and has strong generalization ability. It can also be extended to other machine learning problems such as function fitting.

Transform a straight line that can separate the two types of samples to both sides until it intersects with the sample then two straight lines can be obtained, which are called the Maximum Margin Hyperplane. The wider the classification interval, the better the classification effect. The sample points that fall on the maximum edge hyperplane are called Support Vectors, and the support vector machine gets its name. The idea of classification interval is transformed into mathematical language below. Any straight line in the two-dimensional plane can be expressed in the form of $w_1x_1 + w_2x_2 + b = 0$, abbreviated as $\boldsymbol{x}^T\boldsymbol{w} + b = 0$, where $\boldsymbol{x}$ and $\boldsymbol{w}$ are column vectors$(x_1,x_2)$and(w_1,w_2), T the transpose symbol, b is a constant. The classification hyperplane is expressed as $\boldsymbol{x}^T\boldsymbol{w} + b = 0$, and the largest edge hyperplane above it is correspondingly expressed as $\boldsymbol{x}^T\boldsymbol{w} + b - 1 = 0$, that is $\boldsymbol{x}^T\boldsymbol{w} + b = 1$. Similarly, the largest edge hyperplane below is expressed as $\boldsymbol{x}^T\boldsymbol{w} + b = -1$. The interval between the two largest edge hyperplanes is equal to $2/\|\boldsymbol{w}\|$, where $\|\boldsymbol{w}\|$ is the 2 norm of the vector $\boldsymbol{w}$, that is, the square root of the sum of the squares of the elements.

The goal of a linear support vector machine is to find a set of straight-line parameters $\boldsymbol{w}$ and b, so that the classification interval can be maximized. The objective function can be written as:

$$\max_{\boldsymbol{w},b} \frac{x}{\|\boldsymbol{w}\|}$$

The more commonly used objective function is the equivalent form of the above formula:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|$$

ISSN 2616-5775 Vol. 4, Issue 1: 15-19, DOI: 10.25236/AJCIS.2021.040103

Linear support vector machines can handle linear classification problems, but non-linear classification problems require other solutions. When a straight line cannot be found in the two-dimensional plane to distinguish the two types of samples, the idea of increasing the dimension needs to be introduced. Adding a dimension to the two-dimensional feature and mapping ($x^{(1)}$, $x^{(2)}$) to the three-dimensional feature (($x^{(1)}$)^2, ($x^{(2)}$)^2, $\sqrt{2}x^{(1)}x^{(2)}$) can convert non-linear classification into linear classification. This is the core idea of the kernel support vector machine, which is to first transform the original data into a high-dimensional feature space through nonlinear mapping ϕ, and then use the linear support vector machine to classify the data in the high-dimensional space to solve the non-linear problem [7-8]:

$$\boldsymbol{x} \mapsto (\boldsymbol{x}) = (\phi 1(\boldsymbol{x}),...,\phi k(\boldsymbol{x}),...)$$

After transforming the original data x to high-dimensional data ($\boldsymbol{x}$) through nonlinear mapping $\phi(\boldsymbol{x})$, the objective function of the linear support vector machine dual problem is:

$$\max[\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \, \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)] \quad \boldsymbol{\alpha}$$

The calculation of the kernel function only needs to be performed in the low-dimensional feature space, which greatly reduces the computational complexity.

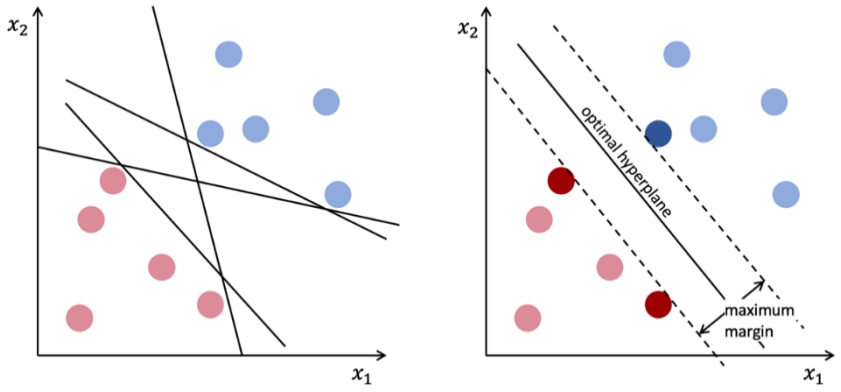

Figure. 1 Illustration of SVM

3. Data

Our data comes from a group of records about Indian liver patient from a website called Kaggle (https://www.kaggle.com). This data collected from North East of Andhra Pradesh, India -- contains 416 patients with liver disease records and 167 healthy patients records, in which includes records of 441 male patients and 142 female patients. The $\boldsymbol{x}$ values of this data set comprise Age of the patient (any patients who is older than 89 years-old is listed as being the age of “90”), Gender of the patient, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio, and Dataset. The dataset field is used to split the data into two sets: patient with liver disease (which we labeled as 1 in the last column), or no disease (which we labeled 2).

Table 1 Statistics of different features

	age	total Bilirubin	direct Bilirubin	total proteins	albumin	A/G ratio	SGPT	SGOT	Alkaline phosphate	1/2
count	583	583	583	583	583	583	583	583	579	583
mean	44.74	3.29	1.48	290.57	80.71	109.91	6.48	3.14	0.94	1.28
std	16.18	6.20	2.80	242.93	182.62	288.91	1.08	0.79	0.31	0.45
min	4	0.4	0.1	63	10	10	2.7	0.9	0.3	1
25%	33	0.8	0.2	175.5	23	25	5.8	2.6	0.7	1
50%	45	1	0.3	208	35	42	6.6	3.1	0.93	1
75%	58	2.6	1.3	298	60.5	87	7.2	3.8	1.1	2
max	90	75	19.7	2110	2000	4929	9.6	5.5	2.8	2

ISSN 2616-5775 Vol. 4, Issue 1: 15-19, DOI: 10.25236/AJCIS.2021.040103

It can be seen from Table 1 that our data comes from the age group of four to ninety years old and above, but most of them come from middle-aged people and the elderly. There are relatively few data on infants and young people, and the data of men is much more than that of women.

4. Results

When a set of data is inseparable in the current dimension, adding dimensions or making some transformations that make it separable is the basic idea of kernel functions. The transformations are called kernels, which popular ones include: Polynomial Kernel, Gaussian Kernel, Radial Basis Function (RBF), Laplace RBF Kernel, Sigmoid Kernel, ANOVA RBF Kernel. Choosing the right kernel is crucial to reliable results. We let the computer try two different kernel function support vector machines, namely Linear and RBF (radial basis function), to calculate the best classification boundary. In the data we collected, 80% of it is used to train the computer to calculate the best classification boundary, and the remaining 20% is used to test the accuracy of the boundary we obtained. The definition of accuracy is

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ samples}.$$

In a binary classification like judging whether a sample is ill. A positive case refers to a sample that is correctly predicted by the model as a positive category (infected), while a negative case is a sample that is correctly predicted by the model as a negative category (not infected). After computer calculations, we get an accuracy rate of about 72.4%.

Table 2 Results of prediction

model	accuracy	precision	recall	f1_score	roc
SVM	0.724138	0.724138	1	0.84	0.5

5. Conclusions

We mainly research and use support vector machines to predict whether a person has liver disease based on a series of data. We obtained a set of data from the Kaggle website that recorded 416 liver disease patients and 167 healthy people in India, including their age, gender, various medical indicators, and disease status. Most of the data are middle-aged and elderly Male. We tried linear and RBF kernel support vector machines with the help of the python language, and calculated the best classification boundary with an accuracy of 72.4% based on the obtained data.

Our model can be further improved to make it have a higher accuracy rate. We can accomplish this by getting more data and ignoring abnormal data. And the calculation results this time may only apply to people in India. Obtaining data from different regions can make our conclusions practical in more places.

References

[1] Li, Yuanzhong, Shoji Hara, and Kazuo Shimura. "A machine learning approach for locating boundaries of liver tumors in ct images." 18th International Conference on Pattern Recognition (ICPR'06). Vol. 1. IEEE, 2006.
[2] Petrea, Stefan Mihai, et al. "A Machine Learning Approach in Analyzing Bioaccumulation of Heavy Metals in Turbot Tissues." Molecules 25.4696 (2020): 1-42.
[3] Minhas, Fu, D. Sabih, and M. Hussain. "Automated Classification of Liver Disorders using Ultrasound Images." (2012).
[4] Cao, Guitao, P. Shi, and B. Hu. "Liver Fibrosis Identification Based on Ultrasound Images." 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference IEEE, 2006.
[5] Comak, Emre, et al. "A new medical decision making system: least square support vector machine (LSSVM) with fuzzy weighting pre-processing." Expert Systems with Applications 32.2 (2007): 409-414.

ISSN 2616-5775 Vol. 4, Issue 1: 15-19, DOI: 10.25236/AJCIS.2021.040103

[6] Tang, Yuchun, et al. "Granular support vector machines for medical binary classification problems." 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2004.
[7] Amari, Shun-ichi, and Si Wu. "Improving support vector machine classifiers by modifying kernel functions." Neural Networks 12.6 (1999): 783-789.
[8] Min, Jae H., and Young-Chan Lee. "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters." Expert systems with applications 28.4 (2005): 603-614.