Liver Disease Prediction using SVM and Naïve Bayes Algorithms

Dr. S. Vijayarani[1], Mr.S.Dhayanand[2]

Abstract— **In recent years in healthcare sectors, data mining became an ease of use for disease prediction. Data mining is the process of dredge up information from the massive datasets or warehouse or other repositories. It is a very challenging task to the researchers to predict the diseases from the voluminous medical databases. To overcome this issue the researchers use data mining techniques such as classification, clustering, association rules and so on. The main objective of this research work is to predict liver diseases using classification algorithms. The algorithms used in this work are Naïve Bayes and support vector machine (SVM). These classifier algorithms are compared based on the performance factors i.e. classification accuracy and execution time. From the experimental results it is observed that the SVM is a better classifier for predict the liver diseases.**

Keywords - **Classification, Liver function test, Naïve bayes, SVM**

I. Introduction

Researchers faces more challenging task in healthcare sectors to predict the diseases from the voluminous medical databases. Nowadays data mining became more essential in healthcare sectors. Data mining techniques which includes classification, clustering, association rule mining for finding frequent patterns are applied to medical data for disease prediction. In data mining, classification techniques are much popular in medical diagnosis and predicting diseases [1]. In this research work, Naïve Bayes and Support Vector Machine (SVM) classifier algorithms are used for liver disease prediction. There are several numbers of liver disorders that required clinical care of the physician [3**]**. The main objective of this research work is to predict liver diseases such as Cirrhosis, Bile Duct, Chronic Hepatitis, Liver Cancer and Acute Hepatitis from Liver Function Test (LFT) dataset using above classification algorithms.

The liver is the second largest internal organ in the human body, playing a major role in metabolism and serving several vital functions, e.g. Decomposition of red blood cells, etc.,. [7**]** Its weight comes around three pounds. The liver performs many essential functions related to digestion, metabolism, immunity, and the storage of nutrients within the body. These functions make the liver as an important organ, without this, body tissues would quickly die from lack of energy and nutrients. There are number of factors which increase the risk of liver disease. Some of them are listed below:

- Family history of liver disease
- Smoking
- Consumption of alcohol
- Intake of contaminated food
- Obesity
- Diabetes

The remaining portion of the paper is organized as follows. Related works are discussed in Section 2. The proposed methodology is given in Section 3. Section 4 analyzes the experimental results. Section 5 gives conclusion..

II. Literature Review

Dhamodharan et.al [3] has predicted three major liver diseases such as Liver cancer, Cirrhosis and Hepatitis with the help of distinct symptoms. They used Naïve Bayes and FT Tree algorithms for disease prediction. Comparison of these two algorithms has been done based on their classification accuracy measure. From the experimental results they concluded the Naïve bayes as the better algorithm which predicted diseases with maximum classification accuracy than the other algorithm.

Rosalina et al [13] predicted a hepatitis prognosis disease using Support Vector machine (SVM) and Wrapper Method. Before classification process they used wrapper methods to remove the noise features. Firstly SVM carried out feature selection to get better accuracy. Features selection were implemented to minimize noisy or irrelevance data. From the experimental results they observed the increased accuracy rate in the clinical lab test cost with minimum execution time. They have achieved the target by combining Wrappers Method and SVM techniques.

Omar S. Soliman et al [10] has proposed a hybrid classification system for HCV diagnosis, using Modified Particle Swarm Optimization algorithm and Least Squares Support Vector Machine (LS-SVM). Feature vectors are extracted using Principle Component Analysis algorithm. As LS-SVM algorithm is sensitive to the changes of values of its parameters, Modified-PSO Algorithm was used to search for the optimal values of LS-SVM parameters in less number of iterations. The proposed system was implemented and evaluated on the benchmark HCV data set from UCI repository of machine learning databases. It was compared with another classification system, which utilized PCA and LS-SVM. From the experimental results the proposed system obtained maximum classification accuracy than the other systems.

Karthik et.al [7] were applied a soft computing technique for intelligent diagnosis of liver disease. They have implemented classification and its type detection in three phases. In the first phase, they classified liver disease using Artificial Neural Network (ANN) classification algorithm.

Dr. S. Vijayarani*, Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India.*

Mr. S.Dhayanand, *M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India.*

ISSN: 2278 – 7798

In the second phase, they generated the classification rules with rough set rule induction using Learn by Example (LEM) algorithm. In the third phase fuzzy rules were applied to identify the types of the liver disease.

Chaitrali S. Dangare et.al [2] has analyzed prediction systems for Heart disease using more number of input attributes. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performances of these techniques are compared, based on accuracy. Authors' analysis shows that out of these three classification models Neural Networks has predicted the heart disease with highest accuracy.

III. Methodology

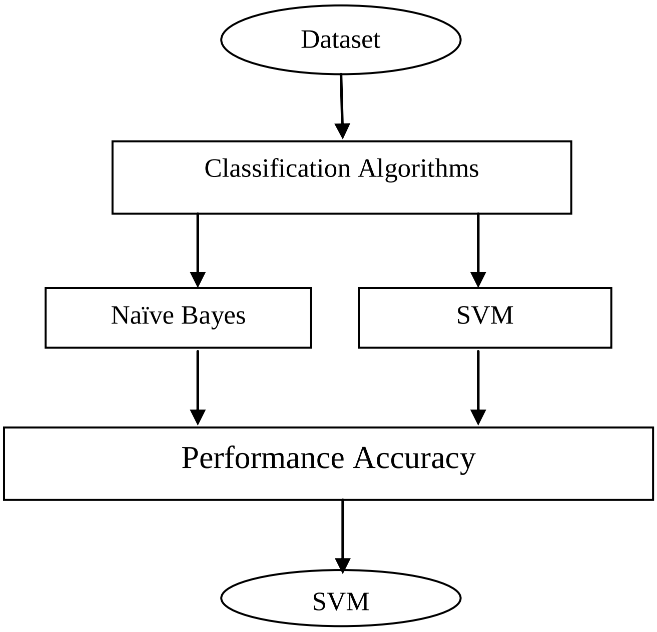

Figure 1: System Architecture

Dataset

Indian Liver Patient Dataset (ILPD) has been taken from the UCI Repository. This dataset has five hundred and seventy six instances and ten attributes. Attributes are Age, Gender, TB, DB, ALP, Sgpt, Sgot, TP, ALB and A/G Ratio. This dataset contains Liver Function Test details (LFT).

Naïve Bayes

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. A more descriptive term for the underlying probability model would be the self-determining feature model. In basic terms, a Naive Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature [11]. The Naive Bayes classifier performs reasonably well even if the underlying assumption is not true.

The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because of independent variables are unspecified, only the variances of the variables for each *label* need to be determined and not the entire covariance matrix. In contrast to the Naive Bayes operator, the Naive Bayes (Kernel) operator can be applied on numerical attributes.

This can be able in a clear-cut fashion using kernel density estimation and Bayes' theorem:

$$\hat{P}(y = j|x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\sum_{k=1}^{k} \hat{\pi}_k \hat{f}_k(x_0)}$$

Where

$\hat{\pi}_j$ is an estimate of the prior probability of class j; usually, $\hat{\pi}_j$ is the sample proportion falling into the j$_{th}$ category

$\hat{f}_j$ is the predictable density at x0 based on a kernel density fit involving only observations from the j$_{th}$ class

This is essentially the same idea as discriminant analysis, only instead of assuming normality, were estimating the probability density of the classes using a nonparametric method Patrick

Support Vector Machine

Support Vector Machine was first found by Vapnik in 1979 [5]. It was again recommended by Vapnik in 1995 for regression and classification [4]. Support vector can be used for pattern classification [8] which has multilayer perceptron and radial-basis function networks [12**]**. The SVM is the advanced technology with maximum classification algorithms embedded in statistical learning theory [9**]**. SVM methods are used in classification of linear and non-linear data. It transforms the original training data into higher dimension using non-linear mapping. Within this new dimension it searches for linear optimal separating hyperplane. Data from two classes can be separated by hyperplane with an appropriate nonlinear mapping to a sufficiently high dimension. Using support vectors and margins the SVM finds these hyperplane [6]. SVM implements the classification task by maximizing the margin classifies both class while minimizing the classification errors. Although the SVM can be applied to various optimization problems such as regression, the classic problem is that of data classification. The basic idea is shown in figure 2. The data points are identified as being positive or negative, and the problem is to find a hyper-plane that separates the data points by a maximal margin.

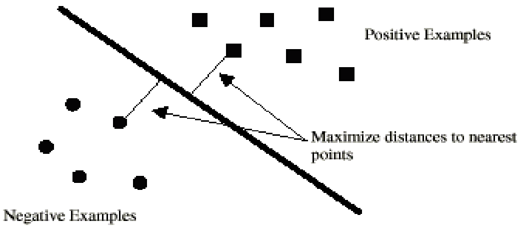

Figure 2: Data Classification

The figure 2 shows the 2-dimensional case where the data points are linearly separable. The mathematics of the problem to be solved is the following:

$$\left.\begin{aligned}
&\min_{\vec{w},b} \frac{1}{2}\|w\|,\\
&s.t \quad y_i = +1 \Rightarrow \vec{w}\cdot\vec{x}_i + b \geq +1\\
&\qquad\; y_i = -1 \Rightarrow \vec{w}\cdot\vec{x}_i - b \leq -1\\
&s.t\ y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1, \quad \forall i
\end{aligned}\right\} \quad (1)$$

The identification of the each data point x_i is y_i, which can take a value of +1 or -1 (representing positive or negative respectively). The solution hyper-plane is the following:

$$u = \vec{w}\cdot\vec{x} + b \qquad (2)$$

The scalar b is also termed the bias.

A standard method to solve this problem is to apply the theory of Lagrange to convert it to a dual Lagrangian problem. The dual problem is the following:

$$\left.\begin{aligned}
&\min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j (\vec{x}_i\cdot\vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i\\
&\sum_{i=1}^{N}\alpha_i y_i = 0\\
&.\\
&\alpha_i \geq 0, \quad \forall i
\end{aligned}\right\} (3)$$

The variables α_i are the Lagrangian multipliers for corresponding data point x_i.

IV. Experimental Results

In this section, the results are analyzed which are given by the classification algorithms such as naïve Bayes and Support Vector Machine. This work is implemented in Matlab 2013 tool.

Figure 3 represents the accuracy measure for the Naïve Bayes and SVM classification algorithms. An experimental result shows the performance of SVM is better than Naïve Bayes algorithm.

Table 1: Accuracy Measure for Liver Disease Dataset

Algorithms	**Correctly Classified Instances (%)**	**Incorrectly Classified Instances (%)**	**TP Rate**	**Precision**	**F Measure**
Naïve Bayes	61.28	38.72	0.612	0.558	0.251
SVM	79.66	20.34	0.796	0.766	0.331

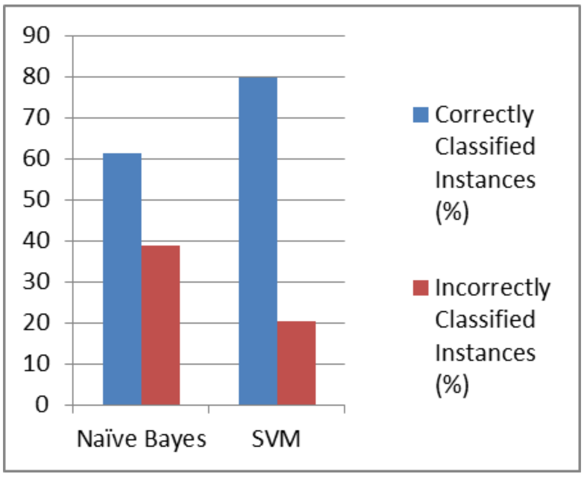

Figure 3: Accuracy Measure

Table 2: Execution time Analysis for Liver Disease Dataset

Algorithms	**Execution Time in ms**
Naïve Bayes	1670.00
SVM	3210.00

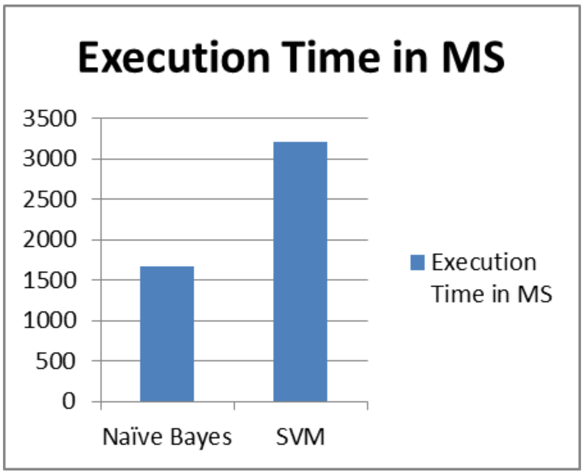

Figure 4: Execution Time Analysis

Table 2 represents the execution time requirement of classification algorithms for predicting liver diseases from liver function test dataset. Figure 4 represents the time taken for execution process. Naïve Bayes performs with minimum period of execution time than SVM.

Table 3: Classification of Liver Diseases

Kidney Disease	Naïve Bayes	SVM
Normal	369	312
CBCL	167	221
Acute Hepatitis	25	26
Outliers	15	17

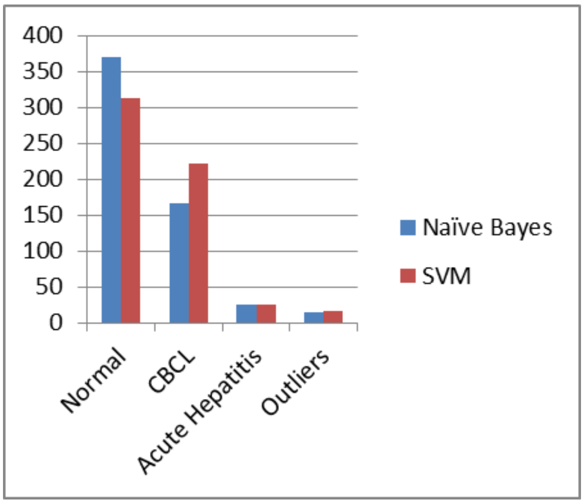

Figure 5: Liver Disease Classification

Table 3 represents and describes the classification of liver diseases such as cirrhosis, bile duct, chronic hepatitis, liver cancer and acute hepatitis. (CBCL) derives first four diseases. CBCL are the liver diseases affected when the liver function test data slightly increased with normal range. Acute hepatitis is a severe liver disease occurs when the liver function test data heavily increases than the normal range. Outliers are predicted in this work based on the moderate range of the liver function test results. Figure 5 represents the classification of liver diseases by classifiers such as Naïve Bayes and SVM. By analyzing the results, SVM gives the overall best classification result than Naïve Bayes classifier.

V. Conclusion

Classification is the major data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. This research work used classification algorithms namely Naïve bayes and Support Vector Machine (SVM) for liver disease prediction. Comparisons of these algorithms are done and it is based on the performance factors classification accuracy and execution time. From the experimental results, this work concludes, the SVM classifier is considered as a best algorithm because of its highest classification accuracy. On the other hand, while comparing the execution time, the Naïve Bayes classifier needs minimum execution time.

REFERENCES

[1] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, *A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis*, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011 page no 101-114

[2] Chaitrali S. Dangare, Sulabha S. Apte, "*Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques*", International Journal of Computer Applications (0975 – 888), Volume 47– No.10, June 2012, page no 44-48

[3] Dhamodharan. S, Liver Disease Prediction Using Bayesian Classification, Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014, page no 1-3.

[4] Grimaldi. M, Cunningham. P, Kokaram. A, *An evaluation of alternative featureselection strategies and ensemble techniques for classifying music*, in: Work-shop in Multimedia Discovery and Mining, ECML/PKDD03, Dubrovnik, Croatia, 2003.

[5] Gur Emre Güraksın, Hüseyin Hakli, Harun U˘guz, *Support vector machines classification based on particle swarmoptimization for bone age determination*, Elsevier publications, Science direct, page no 597-602

[6] Han, J.; Kamber, M., "Data Mining Concepts and Techniques". 2nd Edition, Morgan Kaufmann, San Francisco.

[7] Karthik. S, Priyadarishini. A, Anuradha. J and Tripathi. B. K, *Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types*, Advances in Applied Science Research, 2011, 2 (3): page no 334-345

[8] Kotsiantis. S.B, Increasing the Classification Accuracy of Simple Bayesian Classifier, AIMSA, pp. 198-207, 2004

[9] Milan Kumari, Sunila Godara, *Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction*, International Journal of Computer Sci ence and Technology, Vol. 2, Iss ue 2, June 2011, page no 304-308

[10] Omar S.Soliman, Eman Abo Elhamd, *Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine*, International Journal of Scientific & Engineering Research, Volume 5, Issue 3, March-2014 122

[11] Patrick Breheny, *Kernel density classification*, STA 621: Nonparametric Statistics October 25

[12] Pushpalatha. S, Jagdesh Pandya, *Data model comparison for Hepatitis diagnosis*, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-7) 2014, page no 138-141

[13] Rosalina. A.H, Noraziah. A. Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method, IEEE, (2010), 2209-22

Authors

Dr. S. Vijayarani, MCA, M.Phil, Ph.D working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Mr. S. Dhayanand has completed M.Sc in Software Systems. He is currently pursuing his M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. His fields of research interest are data mining and medical mining. He has published papers in international journals and presented research papers in international, national conferences and Symposiums.