CLASSIFICATION OF CARDIOTOCOGRAMS INTO FETAL STATES USING SUPPORT VECTOR MACHINE

Balagopal Unnikrishnan (A0178398E)
Kenneth Rithvik (A0178448M)
Madan Kumar Manjunath (A0178237W)
Ria Vijay Nagpal (A0178257R)

Master of Technology(MTech.) in Knowledge Engineering (K.E.) /Software Engineering(S.E.)

ABSTRACT

Continuous evaluation, monitoring and correct diagnosis of maternal and fetal health is very important during pregnancy as well as after delivery. Cardiotocography (CTG) is a valuable metric that helps in this process. CTG is the simultaneous recording of Fetal Heart Rate (FHR) as well as Uterine Contractions (UC). There does exist various signal processing methods to programmatically understand the CTG values but the predictive capacity of these methods are still debated upon. Using the patterns and histograms of CTG, doctors can predict or diagnose the current state of the fetus. The data that is labelled by experts in the medical field can later be used to train advanced machine learning algorithms to classify the state of the fetus. In this paper we propose the machine learning technique of Support Vector Machines (SVM) to solve or classify the state of the fetus. This technique gives us a pretty good accuracy when trained with the labelled data that we have. In order to measure the performance of our model, we used metrics such as accuracy, precision, sensitivity/recall, f1-score and ROC curve of each class [3].

1. INTRODUCTION

Support Vector Machine (SVM) is a supervised machine learning technique which performs analysis of data to be classified, maps them as vectors in a multi-dimensional vector space and tries to divide vectors of a particular class or category with as wide a gap as possible [1] & [4]. Many research studies are carried out in SVM and such studies conclude that the performance of SVM model is significantly higher than any data classification algorithm. The various application domain of SVM includes voice recognition, detection of objects, recognition of various types of images, text categorization and data classification [2]. For some of the datasets, SVM performance is highly sensitive based on the values of kernel and cost parameters. Hence, we perform model selection by conducting extensive cross validation and find the most optimal parameters for the model. The only issue associated with model selection and a lot of time is required to find the optimal parameters. The parameters such as kernel functions, standard deviation of the Gaussian kernel, slack variables, weights assigned to the slack variables to help understand the uneven distribution of data, and the total number of examples used for training have an impact on the overall results of the SVM algorithm [3].

2. DATA CLASSIFICATION PROBLEM AND VARIABLE DESCRIPTION

Dataset Source Link:
https://archive.ics.uci.edu/ml/datasets/Cardiotocography

The goal of this classification task is to classify the given dataset into the below three categories:

1. Normal
2. Suspect
3. Pathologic

The input data consists of feature measurements of the fetal heart rate (FHR) and uterine contraction (UC). These features were obtained by processing the cardiotocograms (CTG). Each CTG was then labelled into any of the above mentioned three categories of the fetal states by an experienced doctor. The basic attributes of the dataset are:

LB – Heart rate of fetus (beats/minute)
AC – Count of accelerations (fetus) in a second
FM – Number of times the fetus moves in a second
UC – Amount of uterine contractions in a second
DL – Count of light decelerations (fetus) in a second
DS – Count of severe decelerations (fetus) in a second
DP – Count of prolonged decelerations in a second
ASTV – % of time with unusual short term variability
MSTV – Average of short term variability
ALTV – % of time with unusual long term variability
MLTV – Average of long term variability
Width – width of fetal heart rate histogram
Min – min of fetal heart rate histogram
Max – Max of fetal heart rate histogram
Nmax – Number of fetal heart rate histogram peaks
Nzeros – Number of zeros in fetal heart rate histogram
Mode – mode of fetal heart rate histogram
Mean – Average of fetal heart rate histogram
Median – Median of fetal heart rate histogram
Variance –Variance of fetal heart rate histogram
Tendency –Tendency of fetal heart rate histogram
NSP – fetal state class code (N=normal; S=suspect; P=pathologic)(Target Variable)

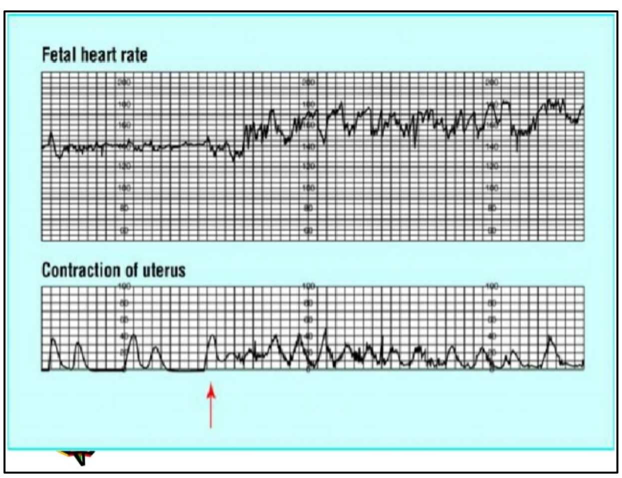

Figure 1: A sample CTG Image

3. DATA PREPROCESSING

Firstly, from the dataset we can see that we have some attributes that can be used only to identify the rows such as id and filenames. There are also some non-predictive attributes like test begin time and test end time. We delete these attributes from the dataset. As the second step, we need to normalize the input variables into values between 0 and 1 using min-max normalization as this greatly helps in the fast computation while building the model as well as not skewing the model towards attributes that have a naturally higher value than others. We also one-hot encode the categorical attribute of tendency into three binary attributes. After this preprocessing is done, we split our dataset into a train and test subset to perform further modeling on. We used a train and test split of 75% and 25% respectively. We have a total of 2127 instances in our dataset which we split into 1595 and 532 test and train datasets respectively.

4. DATA CLASSIFICATION EXPERIMENT RESULTS

First, we perform classification on the dataset using standard baseline SVM model. For such model we use the default values of C, gamma and the linear kernel provided by the library function in the sklearn python package.

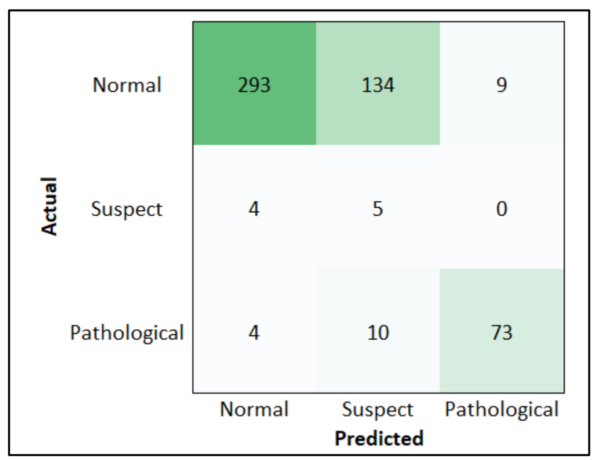

Figure 2: Confusion Matrix for the baseline SVM model

As observed from the confusion matrix, the suspect category was misclassified as normal for most of the cases and hence the model did not classify extensively good. We obtained an accuracy of 69.74% for the baseline SVM Model.

Category	Parameters			
	Precision	Recall	F1-Score	Support
Normal	0.97	0.67	0.80	436
Suspect	0.03	0.56	0.06	9
Pathological	0.89	0.84	0.86	87
Avg/Total	0.94	0.70	0.79	532

Table 1: Classification Report of Baseline SVM with default parameters.

After obtaining a low accuracy using the baseline SVM model, we performed a cross-validated grid-search to obtain optimal values of C and gamma to achieve a model design with high accuracy. We also evaluated the model using different kernel functions and found the radial basis function kernel to be the most optimal one for us. On performing the grid-search we obtained values of C and gamma as 33 and 0.1 respectively. Using these values for the SVM model we again performed classification on the dataset.

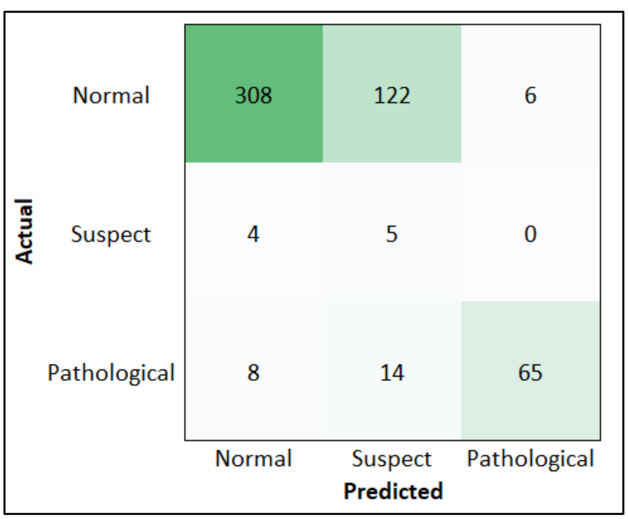

Figure 3: Confusion Matrix for the SVM model with parameters obtained using grid-search

On observing the confusion matrix, we conclude that the suspect category was misclassified as normal for most of the cases and hence the model did not perform its best classification. We obtained an accuracy of 71.05% for this SVM Model designed using parameters found using grid-search.

Category	Parameters			
	Precision	Re-call	F1-Score	Support
Normal	0.96	0.71	0.81	436
Suspect	0.04	0.56	0.07	9
Patho-logical	0.92	0.75	0.82	87
Avg/Total	0.94	0.71	0.80	532

Table 2: Classification Report of SVM Model using grid-search parameters

We also tried other values of C and gamma that we were able to give near optimal solutions according to the grid-search method, we tried to perform classification with the SVM model obtained using these values of C and gamma. The value of C and gamma that we choose is 4.0 and 2.0 respectively.

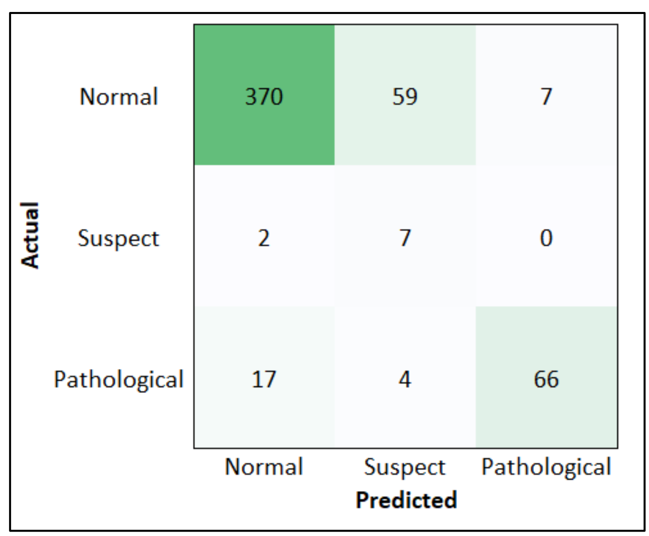

Figure 4: Confusion Matrix for the SVM model with optimal parameter values

As observed from the confusion matrix, the SVM model design with these parameter values has classified the dataset better than the above two methods. The Accuracy obtained for this model design is 83.27%. It should also be noted that for this model the number of pathological cases that were classified as normal is quite more than the previous model. This is the most striking down-side of this model as a false negative for classifying a fetus as pathological might have serious repercussions for the state of the mother as well as the fetus. Therefore we conclude that any model that is able to classify the pathological cases accurately is more valuable and life-saving than the other models.

Category	Parameters			
	Precision	Re-call	F1-Score	Support
Normal	0.95	0.85	0.90	436
Suspect	0.10	0.78	0.18	9
Patho-logical	0.90	0.76	0.82	87
Avg/Total	0.93	0.83	0.87	532

Table 3: Classification Report of SVM Model design with optimal parameters

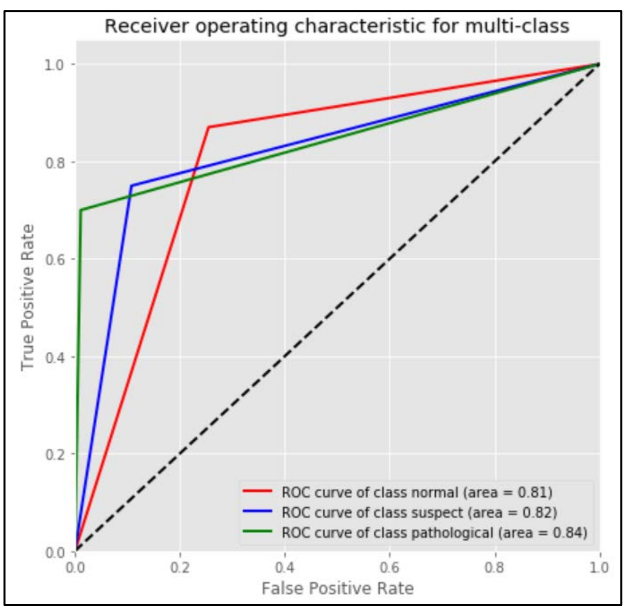

Figure 5: ROC Curve for the SVM Model design

From the ROC curve for all the classes we can see that the data-points are mostly in the upper left hand side of the graph which is desirable as this indicates an higher true positive rate and lower false positive rate.

5. CONCLUSION

We have compared the results of different SVM model designs built using different parameter values of C and gamma. From the results, we can conclude that the for any kernel the selection of values and the function is very crucial and important for any given dataset. As observed from the above results of SVM Models, the optimal parameter values for C and gamma are 4.0 and 2.0 for the cardiotocograms classification dataset as we obtained an accuracy of 83.27%.

6. REFERENCES

[1] Durgesh K. Srivastava, Lekha Bhambhu (2005-2009). Data Classification using Support Vector Machine

[2] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.

[3] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995

[4] https://en.wikipedia.org/wiki/Support_vector_machine