HEB CASS

Support Vector Machine (SVM) Kernels based approach for detection of Breast Cancer

Prity Vijay[1], Bright Keswani[2]

[1]Research Scholar, Department of Computer Sciences, Suresh Gyan Vihar University, Jaipur, India
[2]Associate Professor, Department of Computer Applications, Suresh Gyan Vihar University, Jaipur, India

Email ID- editorcassstudies@gmail.com

ABSTRACT:
Cancer is one of the most dangerous as well as heterogeneous disease .Breast Cancer is the most common forms of Cancer among women. Mammograms, Breast ultrasound, etc. are some of the medical test, commonly prescribed by the doctors, for the diagnosis of breast cancer. But they are not always correct at the beginning stages of breast cancer. Routine checkups are prescribed to every woman after crossing certain age limit and thus exposed to radiation as a side effect of it, increasing cancer risk. Thus there is a need of an alternate solution other than costly and risky medical tests. This paper presents the use of machine learning algorithms for easy reorganization of breast cancer. We build a model using different kernels of Support Vector Machine for the prediction of breast cancer tumor.

Keywords: **Cancer, Mammograms, Breast Ultrasound, Machine Learning, Breast Cancer, Support Vector Machine**

Access this Article Online	Quick Response Code:
Website: http://heb-nic.in/cass-studies	
Received on 03/12/2018	
Accepted on 06/12/2018 © HEB All rights reserved	

INTRODUCTION

The building blocks of our body are cells. These cells combine to form different organs and tissues having different functions. Each of these cells repair and divides themselves in a controlled manner. But in some cases, due to some reason, these cells starts growing in an abnormal fashion and forms lumps which is known as tumor. The overgrowth of cells around the breast duct leads to form breast tumor. The breast tumor can further classified into benign or malignant. Benign tumors are not cancers and can be removed easily without any risk. But malignant tumor, if left untreated the cancer cells can spread further into the breast duct and lobules and hence cause breast cancer. Breast cancer is curable if the symptoms are pragmatic in an early stage but late stages of it are not curable. Among all types of cancer occur in women, breast cancer is the most common one. Approx 1 out of every 8 women in U.S is suffering from breast cancer. Statistics proves almost 25%[1] of the female suffering from cancer is diagnosed with the breast cancer. Although, medical test like mammograms, ultrasounds and others are available but they sometime fails to give 100% accurate result at the beginning of breast tumor. Hence maximizes the chance of growth of cancer cells. Thus creating a lot of risk for physician to predict cancerous tumors correctly. Therefore, a demand for new strategies for the prediction and detection of cancer is needed. Enormous amount of patient big data [2] motivates researchers to apply different machine learning approach and build predictive models, which can accurately predict breast cancer patient.

Machine learning [3] is a field of computer science and was coined in 1959 by Arthur Samuel. Machine learning algorithms are very powerful because its concept is drawn from statistics, mathematics and computer science. Machine learning algorithm learns the rules by applying the concept of statistics and mathematics without any written program for it only through the data provided to it. Machine Learning is further divided into Supervised and Unsupervised. It is consisting of rich library of algorithm among which, support vector machines [4] is one of the most high performing classification algorithm which was introduced in 1992. It is popular because it delivers high performance with little tuning.SVM classifier differentiate classes by creating hyper plane. Hyperplane is a line that splits a plane into two parts where each of two classes lies in either side. The distance between the closest data points and plane is known as margin. There can be many lines separating the classes but the best line is that having largest margin and known as Maximal Margin hyperplane. Hyperplane line for n-dimensional space will be:

$$B_0 + B_1X_1+B_2X_2+ \text{-----} + B_nX_n=0$$

SVM classify the data by maximizing the gap between the support vectors. If the data is linearly separable we can use Maximum Margin Classifier. But in real time scenario data are not always linearly separable. In such cases support vector classifier (SVC) are used to draw hyperplane with some error. Support Vector Machine (SVM), which is an extension to SVC are used for creating non-linear boundaries. Having different set of solution MMC allow us to choose optimal solution .Optimal plane can be selected by calculating the perpendicular distance from every point of a dataset and then selecting the plane with maximum margin.SVM results by transforming feature space in a specific way with the use of kernels. When we deal with high dimensionality data, use of kernels are must in order to give high performance. SVM classifier as a combination of support vector classifier and non linear kernels. Among various types of kernels we can select best kernels depending on the dataset. Apart from using different types of kernels we can tune the classifier by changing the value of coefficient associated with these kernels in order to reach best classification. In this paper we will use different SVM kernel to build a predictive model which can accurately predict the breast cancer tumor without any confusion.

I. RELATED WORK

Data mining and Machine Learning is consist of many techniques which have been using in the medical field from long back. Lots of research has been conducted using different dataset to predict cancer. Hiba Asri [5] compares the performance of different classification ML algorithm. In the experiment, he compares SVM, NB, KNN and c4.5 in terms of accuracy, precision, sensitivity and specificity to find the best accurate result and concluded SVM as a best performer. Abdullah-Al Nahid and Yanan Kong [6], elaborately explains involvement of machine learning in the classification of image for breast cancer. They made an effort to present detail discussion of Convolutional Neural Network (CNN) method for breast image classification. Abdel-ilah L. and sahinbegovic H.[7] uses Artificial Neural Network in order to design a model which and concluded , ANN performances with number of neuron in hidden layers, with best network design can be configured with 3 hidden layers and 21 neurons in the hidden layer with TANSIG activation function. In a paper presented by Kihan Park and javdev p.desai [8] breast cancer is identified using soft-margin SVM. Soft margin SVM separate non-linear separable dataset. They uses custom built microscope compatible micro indentation system where each indentent is label as normal or cancerous according to the image certified by pathologists on which SVM is applied to categorize among these two point. The experiment gives better result when single parameter is took under consideration. Apart from these there are many papers where we can find impact of machine learning for the classification of breast cancer.

II.DESCRIPTION OF BREAST CANCER DATASET

In this paper we are using the dataset obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg[9]. The name of this dataset is Wisconsin Breast Cancer Database which is freely available in internet in a site known as UCI. This dataset contains the record of 699 patients consisting altogether 11 attributes. Our aim is to classify Label_class attribute of this dataset consisting two types of tumor (benign or malignant).As discussed earlier benign is non dangerous tumor while malignant can cause cancer and hence should be predictable for further investigation. This dataset uses 2 for symptomatic of benign, 4 for symptomatic of malignant. Dataset consist 16 missing values, removing them is not a good option as we can lose valuable information, we had replaced all missing values with standard number format -99999.

III. TOOLS USED IN EXPERIMENT

This experiment is performed in a system having a Intel® Core(TM) i7 -3632QM CPU @ 2.20GHz 2.19 GHz with 8 GB RAM and 64-bit Operating System. In this paper we used Jupiter notebook version 3.6.4 as a software environment. For performing machine learning operation we used Scikit-learn as it contains rich sets of Machine learning library. It is available freely and designed to be compatible with python numerical and scientific library NumPy and SciPy. Finally, for plotting graph we have used Matplotlib, which is a visualization tool for python and NumPy.

IV. DATASET OBSERVATION

The Wisconsin Breast Cancer Database is consisting of 699 records and 11 attributes. The first attribute in the dataset is id had been dropped to form a new dataset with 10 attributes. Other 9 attributes of the dataset contains random values ranging between 1 to 10. As presented in figure 1, we can see that we have 458 cases of benign and 241 cases of malignant.

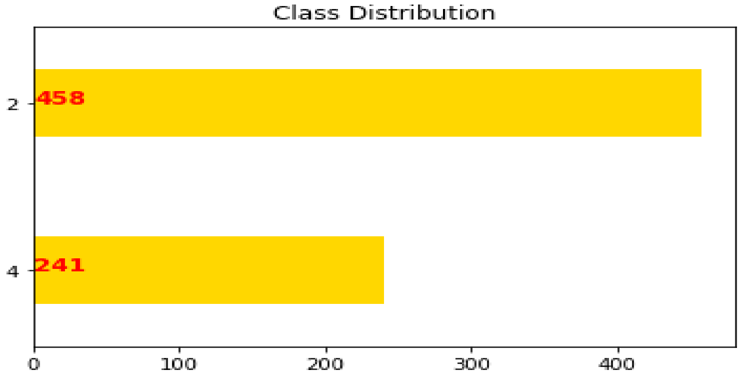

Fig. 1 : Distribution of the Class of the Breast Cancer Data Set

To have more clear idea about dataset we find the density of each attributes in the dataset(see figure 2).Density plot is normalized form of distribution of each attribute of dataset from where we can have an idea how densely the value of each of these instance is distributed. For example, from this graph we can say that almost 90% value of attribute mitosis range between 1 to 2.

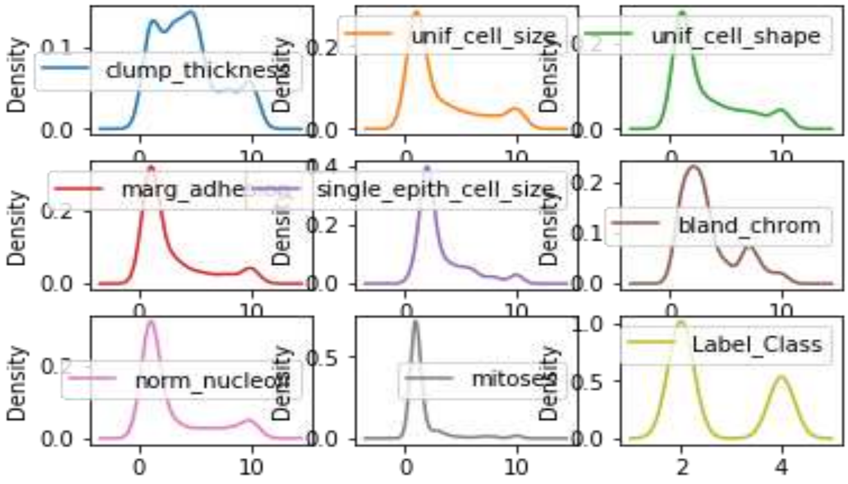

Fig 2 : Density graph for the attributes of the Breast Cancer Dataset.

In order to observe correlation among the attributes of a dataset we have constructed heatmap plot. In the figure 3, we can see that dataset are presented in the form of matrix where each square in the matrix denotes the correlation among two attributes. Lighter shade of square in a graph represents strong correlation and darker shade denotes weak correlation. Here we can observe that mitosis is the only attribute which is weakly correlated with all other attributes. On the other hand Label_class is highly correlated to unif_cell_size and unif_cell_shape. So with the help of this graph we can say that these two variable are very important attributes for the experiment to predict class.

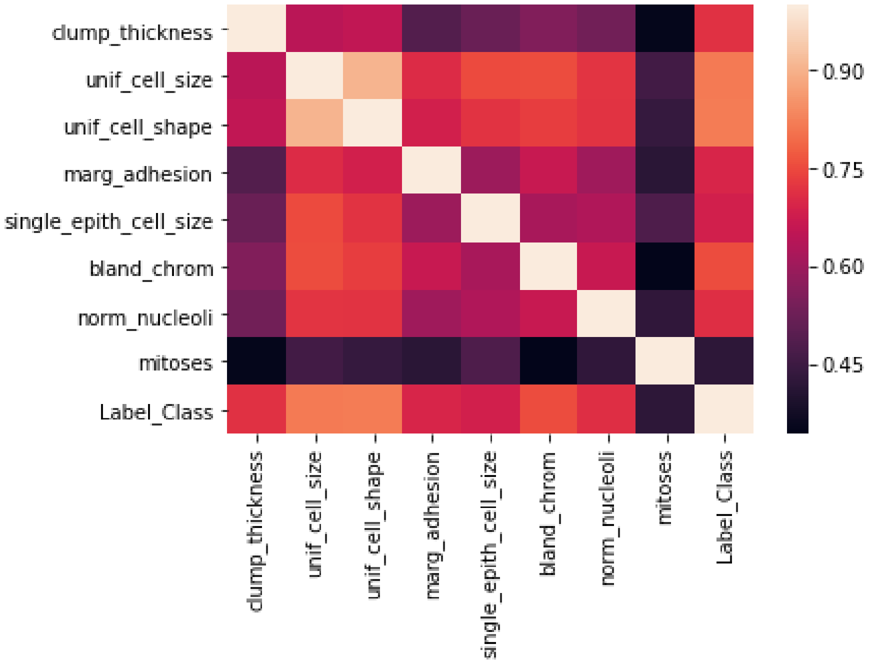

Fig. 3: Heatmap plot to determine the correlation among instances of breast cancer dataset

V. EXPERIMENT RESULT

Our main motive is to build an accurate predictive model for the prediction of breast cancer tumor. We divided the dataset randomly into two parts, training dataset (80% of dataset) and test dataset (20% of dataset).We build a model using support vector machine. We check the model with different types of kernels used in SVM. There after we tuned the parameters by using different values for C and gamma co-efficient. In an experiment (See Table 1) we compares model with 3 different SVM kernels. The performance of poly kernel was worst, therefore we stop considering it. We have seen that linear kernel with gamma value 1 and c value 1000 has given the accuracy of 98% which is the best among all. We have noticed that accuracy is more when the value of gamma is lower and value of C is higher. We mentioned in a table the parameters values which have shown best result.

VI. ALGORITHM TO DEVELOP SVM MODEL

```
Dataset =X
Split  X
X_training _size=0.8
X_test_size=0.2

Compute SVM classifier by using different SVM kernels to draw best hyperplane
Store the result in variable model
Train the model with X_training_size and X_test_size
Check accuracy
```

Table I.
SVM MODEL WITH DIFFERENT KERNELS ALONG WITH DIFFERENT VALUES FOR GAMMA AND C PARAMETER

Kernel	**Accuracy(%)**	**Gamma**	**C**
Linear	95	1	1
	96	1	10
	97	1	100
	98	1	1000
RBF	80	1	1
	82	1	10
	87	1	100
	87	1	100
Po	63	1	1

l y			

CONCLUSION

Breast cancer commonly occur in women mostly after the age of 35.Maximum percentage of patient suffering from this disease dies if it is not known at the beginning. As the risk is high we need a solution with best result. For this purpose we use SVM along with linear, RBF and poly kernels to build predictive model. We observed that linear and RBF kernel performs really very well and poly kernels were worst. We further tuned the model by changing gamma and C parameters. Lastly, we conclude SVM model with linear kernel (gamma=1, c=1000) with an accuracy of 98% as a best model for predicting breast cancer.

REFERENCES

[1] Vikas Chaurasia, Saurabh Pal & etal, “Prediction of benign and malignant breast cancer using data mining techniques”, Journal of Algorithm & Computational Technology,2018

[2] Prity Vijay, Bright Keshwani,” Emergence of Big Data with Hadoop : A Review”,IOSR Journal of Engineering (IOSRJEN)”,Vol:06,Issue:03,2016,pp50-54

[3] https://en.wikipedia.org/wiki/Machine_learning

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

[5] HibaAsria, Hajar Mousannif & etal,“Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis”, The 6th International Symposium on Frontiers in Ambient and Mobile Systems(Elsevier), Vol:83, 2016, pp:64-69

[6] Abdullah-Al Nahid and Yinan Kong,“ Involvement of Machine Learning for Breast Cancer Image Classification: A Survey”, Computational and Mathematical Methods in Medicine, 2017

[7] Layla Abdel-Ilah, Hana Šahinbegović ,“Using Machine Learning Tool In Classification Of Breast Cancer” **,** Springer,Vol:62,2017

[8] Kihan Park ,Jaydev P. Desai “Machine learning approach for breast cancer localization” IEEE Explorer,2017

[9] Breast Cancer Wisconsin (Original) Data Set https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29