Breast Cancer Prediction

S.Pranave, N.Harika, C.Indu

Abstract - In this project, we analyse the existing data on breast cancer and build a model that tells us if the given sample of lump is harmful(Malignant) or harmless(Benign)

I Introduction:

Breast Cancer is one of the major causes of death among women in the modern day.Thus the need for predicting the behavior of the lumps, whether it is malignant or benign is of high importance and is what drives us towards this project.

II Data Description:

Data Set Source:

This paper used a breast cancer database from University of Wisconsin Madison, Clinical
Sciences Center.
The database has about 690 instances and data of the attributes obtained from FNA(fine needle aspirate) procedure:

1. First drawing the aspirate
2. Staining the cells
3. Enhancing their boundaries using computer curve program.

The database uses ten different attributes topredict two fields B=benign or M=malignant.

Wisconsin Dataset :

The data used has total 699 instances, 10 attributes plus the class attribute:

1. Sample code number,
2. Clump Thickness,
3. Uniformity of Cell Size,
4. Uniformity of Cell Shape,
5. Marginal Adhesion,
6. Single Epithelial Cell Size,
7. Bare Nuclei,
8. Bland Chromatin,
9. Normal Nucleoli,
10. Mitoses,
11. Class.

And the class distribution is

1. Benign: 458 (65.5%) and
2. Malignant: 241 (34.5%).

Preprocessing :

There were unfilled portions in our data set. The missing data in the dataset were replaced with the average of its neighbours.

III Building Models:

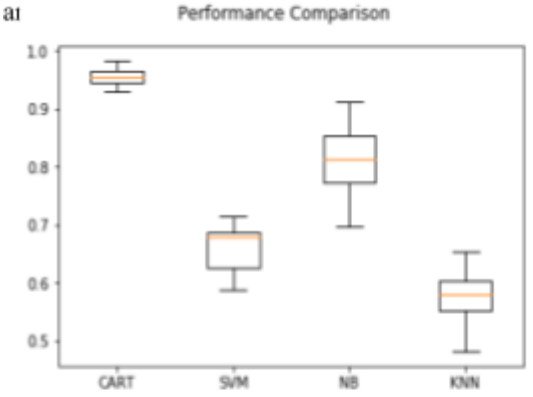
Major Techniques used

1. Support Vector Machine
2. Naive Bayes

Initially in order to choose what models/data mining techniques best suit our requirements, we analysed the performance of four classification models, i.e.

1. Decision Tree
2. Gaussian Naive Bayes
3. Support Vector Machine
4. K-Nearest Neighbours Classifiers

The following are the results of the above mentioned al

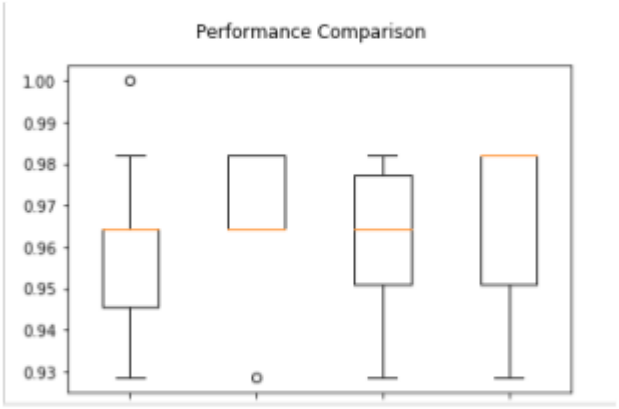


Accuracy, Error rate and run time of various models are

```
CART: 0.953474 (0.016411) (run time: 0.025088)
SVM: 0.663734 (0.043218) (run time: 0.252720)
NB: 0.808701 (0.062160) (run time: 0.025172)
KNN: 0.576169 (0.046208) (run time: 0.026724)
```

Even though according to the above obtained results, SVM's performance isn't up to the mark, we found that after normalizing the data set and doing the same analysis, SVM gives the best result. This is because Support Vector Machine works better with normalized data as compared to non-normalized data.

The following are the results of the analysis after normalizing the data.



Accuracy, Error rate and run time of various models are shown below

```
ScaledCART: 0.958831 (0.021264) (run time: 0.034498)
ScaledSVM: 0.967825 (0.015538) (run time: 0.053770)
ScaledNB: 0.960682 (0.019197) (run time: 0.031538)
ScaledKNN: 0.967825 (0.019209) (run time: 0.043032)
```

As shown above, both SVM and Naive Bayes have good accuracy and less error rate and therefore we intend to work with both.

We have analysed the SVM with different values of c and kernel and found out that c = 0.1 and kernel = linear gives good accuracy rate.

Also Since Breast cancer is a health related sensitive topic, one of our major concerns is to avoid false negatives.That is we ought to avoid telling a person who has cancer that he doesn't have cancer. That way we can avoid increasing the risk of not providing treatment to a person who has cancer.

We ensure this by using two data mining models (SVM and Naive Bayes) and concluding the result as positive even if one of the models says positive.

Results:

Accuracy score of the built model 0.957143
Percent of False Negatives: 0.4149377593360996

Future Work:

We have to analyse the algorithms to know how they are reducing the error rates (specifically false negatives) and come up with an algorithm that further reduces the results.

References:

[1] 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics
Runjie Shen and Yuanyuan Yang Fengfeng Shao
Intelligent Breast Cancer Prediction Model Using Data Mining Techniques

[2] Xiangchun Xiong ,Yangon Kim , Yuncheol Baek, Dae Wong Rhee and Soo-Hong Kim Analysis of Breast Cancer Using Data Mining & Statistical Techniques

[3] Breast Cancer Wisconsin (Diagnostic) Data Set
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)