Training and Testing Classification Models for Breast Cancer Diagnosis

By Neel Paratkar

Abstract

Breast Cancer is one of the most common types of cancer. It is the second leading cause of death in women which contributes about 3.8 percent deaths around the world. Breast Cancer Diagnosis and Prognosis is one of the largest fields of study in medical field. In recent decades, a trend of computer aided breast cancer diagnosis has gained traction due to the reseach and development in the field of Artificial Intelligence and Machine Learning. Fine Needle Aspiration is one of the diagnostic procedures used to investigate the tumors that develop in breast tissues. The sampling thus done can be used to gather data which can be used to train Machine learning Models. The main goal of this project is to train and compare different models using the WDBC dataset which help in classifying the data into "Malignant" and "Benign" categories.

Introduction

Breast Cancer is one of the leading cause of deaths among women. Early detection and treatment of breast cancer is the only possible way to avoid possibly fatal symptoms. Fine Needle Aspiration is one of the most famous diagnostic procedures used to investigate the breast tumors. The sampling thus done can be used to gather data which can be used to train Machine Learning Models which in turn can help in easy diagnosis. One such datasets available for training models is Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The creators of this dataset are Dr. Wolberg of General Surgery Department and W. Street and Olvi Mangasarian of Computer Science Department of University of Wisconsin, Madison. This is one of the most popular datasets which has been used to train classification models for cancer diagnosis.

About the dataset

- The features extracted from the sampling done by FNA are computer from digitized image of FNA of breast mass. These features describe the characteristics of the cellular nuclei present in the images.
- The number of instances present in this dataset are 569. The data set has two main classes "B" for Benign and "M" for Malignant.
- There are "357" Benign entries and "212" Malignant entries.
- There attributes in the data are as such : ID number, class ("M" or "B"), followed by 30 real valued features.
- The ten characteristics of the cell nuclei that are observed are:
  - Radius (mean distance from center to perimeter)
  - Texture
  - Perimeter
  - Area
  - Smoothness
  - Compactness
  - Concavity
  - Concave points
  - Symmetry
  - Fractal Dimensions
- For each characteristic through multiple sampling mean, standard error and largest value are noted. Hence there are 30 features (3 x 10)

Goals of the project

The main goal of the project is to use the WDBC dataset to train classification models that can classify the entered data into either one of the two classes "B" or "M". This project has been implemented using the sci-kit learn library which is one of the best machine learning libraries implemented using Python. The classification models that have been trained and compared are Naive Bayes, Support Vector Machines and Linear Discriminant Analysis. The performace of these models has been measured in terms of the prediction Accuracy and Precision, over multiple iterations differing in the number of features used to train the models. The feature selection has been carried out by using three methods, viz. Feature selection using Principal Component Analysis, Feature selection using Correlation Values and Recursive Feature Elimination method.

Classification Models

Naive Bayes

Naive Bayes is a family of probabilistic classifiers which depend upon the independence of the features of the data being classified. Naive Bayes based models are often used in medical diagnosis systems. It is one of the highly scalable

classifiers and can be trained relatively quicker than other complex classification models.

Support Vector Machines

Support Vector Machine is a family of supervised classifiers. These can be efficiently used in case of binary classification of training datasets. SVMs support a variety of non-linear classification options that use "kernel tricks". One such kernel trick used in this project is "RBF" which stands for Radial Basis Function kernel.

Linear Discriminant Analysis

Linear Disciminant Analysis or LDA is a classification methods used often in machine learning and natural language processing. This type of classification works by trying to represent one feature as a linear dependency function of other features in the training dataset. This representation thus can be helped to find out which features best describe the data and hence help in better classifying the test data.

Feature Selection

Feature Selection is the process of identifying and selecting a subset of input features that can be best used to describe the data. The main idea behind pruning the number of features that are used for training models is that, in many cases, the data contains features that are highly co-related which can introduce redundant information and thus create a harmful bias in classification. Following are few of the objectives as to why feature selection is used:

- Reducing number of features can reduce the time required to train the models
- To avoid the complications that arise due to high number of dimensions. High dimensional models often tend to be complicated and are not easily represented or comprehended.
- It makes interpreting the classification model easy for the analyst
- Help in reducing the amount of overfitting

Principal Component Analysis

Principal Component Analysis is a procedure that is used to reduce the dimensionality of the data. It is a method of transformation of the data which contains highly co-related values into a representation that contains highly independent values. These values are usually orthogonal vectors that are formed using linear composition of the features. These independent values are called principal components. The three classification models Naive Bayes, SVM and LDA have been trained using data with reduced dimensions which resulted from selecting n-best principal components. The tables in figures 1, 2 and 3 the accurary and precision of the models trained using specified number of principal components. From the data we can see that the Naive Bayes model and SVM model are not affected for this dataset by feature selection through PCA. However, LDA performs extremely well particularly when number of PCs selected is 20, with accurary of 96.3 percent and presicion of 97.1 percent.

Model : Naive Bayes

No. of PCs	Accuracy	Precision
10	0.689	0.669
15	0.689	0.669
20	0.689	0.669
25	0.689	0.669
30	0.689	0.669

Figure 1: Performance of Naive Bayes model under feature selection by PCA

Model : SVM

No. of PCs	Accuracy	Precision
10	0.894	0.935
15	0.894	0.935
20	0.894	0.935
25	0.894	0.935
30	0.894	0.935

Figure 2: Performance of SVM model under feature selection by PCA

Recursive Feature Elimination

Recursive Feature Elimination is the method of feature selection that iterates multiple times over the dataset training the model recursively with smaller subset of the features. The model first starts with all the features and thus assigns weights to each feature. Based on these weights the features are selected for next iteration. At the end of the iterations, we are left with a pre-mentioned number of features that best describe the data. Below tables in figures 4,5 and 6 tabulate the performance of the models when trained by pruning the number of features using RFE. The number of features selected is increased from 5 to 25.

We can observe from the tables that Naive Bayes model performs poorly under this method of feature selection. SVM and LDA models perform very well. SVM has highest accuracy and precision when 5 features are selected, while, LDA has highest accuracy and precision when 25 features are selected.

Feature Selection based on Correlation Values

Often it so happens that the data contains many values that are highly correlated to each other. These values may introduce a bias in the classification model that can lead to bad

Model : LDA

No. of PCs	Accuracy	Precision
10	0.942	0.947
15	0.956	0.963
20	0.963	0.971
25	0.96	0.967
30	0.965	0.971

Figure 3: Performance of LDA model under feature selection by PCA

Model : Naive Bayes

No. of features	Accuracy	Precision
5	0.627	0.314
10	0.627	0.314
15	0.629	0.814
20	0.733	0.836
25	0.828	0.839

Figure 4: Performance of Naive Bayes model under feature selection by RFE

Model : SVM

No. of features	Accuracy	Precision
5	0.946	0.928
10	0.921	0.894
15	0.894	0.935
20	0.889	0.92
25	0.886	0.913

Figure 5: Performance of SVM model under feature selection by RFE

Model : LDA

No. of features	Accuracy	Precision
5	0.942	0.95
10	0.949	0.955
15	0.942	0.953
20	0.953	0.963
25	0.963	0.971

Figure 6: Performance of LDA model under feature selection by RFE

performance. We can use the correlation matrix to help identify the features that are highly correlated. Figure 7 shows the correlation matrix of the 30 features present in the data. Tables in figure 8, 9 and 10 show the perfomance of the models when the feature selection is done based on correlation values. From the tables we can see that LDA performs extremely well under this scheme with a maximum accuracy of 96.3 percent and maximum precision of 96.9 percent, when the features with correlation value more than 0.85 (and 0.95) are pruned.

Observations

From the performance matrices above we can see that LDA performs significantly better than the other two models for all three methods of feature selection. SVM model performance is a close second. LDA is a classification model that assumes that the data that is used for training is distributed normally i.e. there is identical distribution of data. SVM on the other hand assumes that all classes of the data are seperable. LDA uses linear functions to map the data while the SVM used in this project uses non-linear RBF kernel function. For this dataset, we can clearly see that LDA performs consistently well as the feature selection in all three methods is performed primarily based on linear correlation of the features. It can also be observed that LDA performace increases as the number of linearly dependent feature increases(i.e. more features are selected) while SVM model has appropriately opposite trend.

Conclusion and Learning Experience

In this project I have tried to implement a few of the models that I personally understood better than other classifier types. Working on this project helped me revise the entire syllabus taught in the Data mining and Artificial Intelligence courses. It provided a better understanding of how intelligent systems can be designed (at a very fundamental level) to analyse the data. Futher implementation can also be done by hyperparamter optimization and training better models using genetic algorithms. I also experimented by training deep learning models however the due to the small size of the data set, the models get trained as well as I would have hoped for.

References and Sources

- Hiba Asria ,Hajar Mousannifb ,Hassan Al Moatassime ,Thomas NoeldUsing , Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, 6th International Symposium on Frontiers in Ambient and Mobile

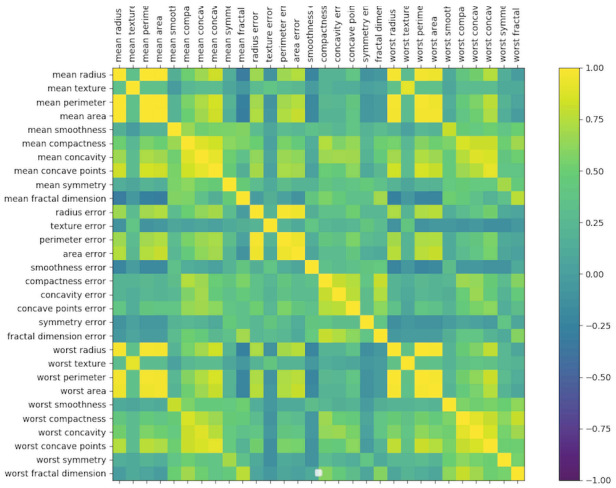

Figure 7: Correlation Matrix of features

Model : Naive Bayes

No. of features	Accuracy	Precision
0.7 >	0.703	0.75
0.75 >	0.719	0.771
0.8 >	0.719	0.771
0.85 >	0.719	0.771
0.9 >	0.775	0.822
0.95 >	0.77	0.785

Figure 8: Performance of Naive Bayes model under feature selection by Correlation Values

Systems (FAMS 2016), Procedia Computer Science 83 (2016) 1064 1069

- Tingting Mu, Asoke K. Nand, BREAST CANCER DIAGNOSIS FROM FINE-NEEDLE ASPIRATION USING SUPERVISED COMPACT HYPERSPHERES AND ESTABLISHMENT OF CONFIDENCE OF MALIGNANCY, 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008
- Breast Cancer Wisconsin (Diagnostic) Data Set, archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names
- Online Resources:
  - BreastCancer.org : www.breastcancer.org
  - Sci-kitLearn.org : scikit-learn.org
  - Wikipedia: www.wikipedia.org
  - Stack Overflow: stackoverflow.com

Model : SVM

No. of features	Accuracy	Precision
0.7 >	0.91	0.885
0.75 >	0.908	0.881
0.8 >	0.91	0.884
0.85 >	0.91	0.884
0.9 >	0.921	0.895
0.95 >	0.929	0.904

Figure 9: Performance of SVM model under feature selection by Correlation Values

Model : LDA

Correlation	Accuracy	Precision
0.7 >	0.954	0.96
0.75 >	0.958	0.964
0.8 >	0.963	0.968
0.85 >	0.963	0.969
0.9 >	0.96	0.968
0.95 >	0.963	0.969

Figure 10: Performance of LDA model under feature selection by Correlation Values