Report for Classification of Breast Cancer Cells

Author: Alperen Demirci - Mail: alperendemirci65@gmail.com

Problem Definition

What is the problem?

We're given a bunch of different patients which have or don't have breast cancer. Our aim is to predict whether if the patient has breast cancer or not. The data is obtained by mammograms of patients. Depending on the data mining done to mammograms, we can create different features from a single photo. That's why there are similar but different datasets on Machine Learning repository.

Inspecting other researches and machine learning projects done on this problem, I can say that they are pretty successful since most of them have F1 Score and Accuracy over 0.9 . The key part in here is Data Preprocessing. Hence features do not have a normal or known distribution, it's hard to process the data. Hopefully we can cluster the data into two clusters using Means with PCA. This helps us to diversify the entries even though we don't have a normal like distribution.

When we check the correlation between variables (including the target variable), there exist a high correlation between nearly every variable. This is both beneficial and harmful to our aim. High correlation between target and features helps us to predict the target variable, whilst high correlation between predictors lead to increment in model complexity and overfitting.

We will deal with the top 3 most used datasets which are: Wisconsin Breast Cancer Diagnosis, Wisconsin Breast Cancer Original, Breast Cancer Dataset. These datasets are created for different aims. Some of them are gathered just for statistical inference, whilst others are created to use in Machine Learning. We will mainly focus on the ones which are created for Machine Learning. Also we should consider that these data are realistic, trustable and have a good quality.

I've compared these 3 datasets in the Jupyter Notebook but in case it may be hard to focus, I'll write the same things here.

Dataset Selection

* We have three dataset options for this task.
  * Breast Cancer Dataset (https://archive.ics.uci.edu/dataset/14/breast+cancer)
  * Breast Cancer Wisconsin (Diagnostic) Data Set (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)
  * Breast Cancer Wisconsin (Original) Data Set (https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original)

Our choice will be based on different parameters such as number of features, number of entries, and the quality of the dataset.

We will compare the two datasets and select the most suitable one for our task.

Breast Cancer Dataset

- Number of features : 9
- Number of entries : 286
- Quality of the dataset : 7.5/10
- Citation Number: 147
- View Count: 107940
- Has missing values : Yes
- Info: This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Breast Cancer Wisconsin (Diagnostic) Data Set

- Number of features : 30
- Number of entries : 569
- Quality of the dataset : 9/10
- Citation Number: 37
- View Count: 282759
- Has missing values : No
- Info: Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Breast Cancer Wisconsin (Original) Data Set

- Number of features : 10
- Number of entries : 699
- Quality of the dataset : 8/10
- Citation Number: 6
- View Count: 101257
- Has missing values : Yes
- Info: This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Features

Since it takes too much space to write all the features here, I will only write the comment about the features.

- Breast Cancer Wisconsin (Diagnostic) Data Set features are all **continuous variables** which helps us to apply the classification algorithms. Also, it has 30 features which is a good number for the classification task. When we check the distribution for the features from Kaggle, we can see that **the features are normally distributed** which is a good sign for the classification task.
- Breast Cancer Dataset features are **multivariate** which means we need to work on preprocessing more than Wisconsin dataset. Also, it has 9 features which is not bad for a classification task. Negative side of the dataset is that all **continuous variables are binned** which **loses information** about the dataset.
- Breast Cancer Wisconsin (Original) Data Set features are all **integers**. Also, it has 10 features which is not bad for a classification task. Negative side of the dataset is that all variables are **compressed** in an integer data type instead of float which **loses information** about the dataset.
- However, the dataset is obtained from the same source with the Wisconsin (Diagnostic) Data Set with a different approach. Therefore, we can say that the quality of the dataset is good.

Conclusion

- The provided dataset on the assignment pdf is Breast Cancer Wisconsin (Original) Data Set. Comparing this dataset with the other two, we can say that the quality of the dataset is good and it has a good number of features for the classification task. However, in my opinion the best dataset for the classification task is Breast Cancer Wisconsin (Diagnostic) Data Set. It has the most number of features and the quality of the dataset is the best among the three. Although it has the most number of features, it does not have missing values(unlike Breast Cancer Wisconsin (Original)) which is a good sign for the dataset.

- If I rank these 3 datasets, I would rank them as follows:
  1. Breast Cancer Wisconsin (Diagnostic) Dataset
  2. Breast Cancer Wisconsin (Original) Dataset
  3. Breast Cancer Dataset

- Since the Original dataset is provided in the pdf and the quality of the dataset is good, I will use the Original dataset for the classification task. If I had to make a choice, I would choose the Wisconsin (Diagnostic) Data Set since it's widely used in Machine Learning Community and requires less preprocessing.

Explanatory Data Analysis (EDA)

1. *Data Preprocessing*
   * *Missing Values* (Found 16, Dropped all of them)
   * *Outliers* (Applied Log Transformation since features are right skewed.)
   * *Feature Scaling* (Used Robust Scaler since it's robust to outliers.)
   * *Encoding* (Skipped since Robust Scaler handled the encoding for target variable.)

2. *Exploratory Data Analysis*
   * *Correlation* (Plotted the correlation matrix.Applied the feature selection.)
   * *Distribution of the Features* (Inspected the distributions of features.)
   * *Clustering and PCA* (Plotted the data using PCA's first two components. Clustered it into two clusters using KMeans and used clusters as a predictor.)

3. *Model Building* (Trained models with different algorithms.)
   * *Train Test Split* (Split the data for training and testing using stratifying due to class imbalance.
   * *Support Vector Classifier*
   * *Random Forest Classifier*
   * *Logistic Regression Classifier*
   * *XGBoost Classifier*
   * *Neural Network Classifier*

4. *Model Evaluation* (We will talk about this in report.)

5. *Conclusion*

Classification Performances

Model	Accuracy	Presicion	Recall	ROC-AUC Score
Logistic Regression	0.9708	0.9230	1.0000	0.9775
kNN Classifier	0.9562	0.9200	0.9583	0.9567
Random Forest	0.9635	0.9215	0.9792	0.9671
Support Vector Machines	0.9635	0.9215	0.9792	0.9671
Neural Network	0.9781	0.9411	1.0000	0.9831
XGBoost	0.9708	0.9230	1.0000	0.9775

Note: Results may differ slightly due to the restart of the python kernel.

Conclusion and Evaluation

- Since we are dealing with a disease problem, our main concern should be **precision**. We can give another reason as there exist a class imbalance and we should consider it when interpreting the results. There are two main and important reasons for why we should check precision at the first place. To sum up, we don't want to classify a diseased person as healthy.

- When we check the values for precision we can see that **Neural Networks** has done the best job among other models. After Neural Networks, **Logistic Regression and XGBoost** algorithms performed well in the second place.

- Also when we check the ROC-AUC Score which is an important metric since it handles both Precision and Recall, we can see that **Neural Networks** is again **the best one.**

- The **worst** algorithm in this table is my favorite classifier: **KNN**. Even if I tuned it with Grid Search it still gave the worst result among others. I think it's because the power of **Neural Networks and XGBoost** in finding the **complex and nonlinear** relations in data are more powerful than the **similarity between entries with same class.**

- **Random Forest** and **SVM** has given the same results so far. I don't think that's because of the models nature. In my opinion the reason is in data itself. It's small sized so you can not see the little nuances between two different models by classification results. Overall, they both performed good but not the best.

In conclusion, I'd choose Neural Networks over any models since it's better than every other model in metrics. Especially **precision** and **roc-arc score** values are important in this dataset, so we have a winner.

- **Winner = Neural Networks (Baba…)**

Thanks for reading!
Please mail me if you've noticed any mistake or have a suggestion!