REPORT

Introduction to Focus Areas

Group 3: Dennis Kragen, Ibraim Ibraimi, Melika Moradi, Neeraj Chauhan, Shipra Guin

Full list of author information is available at the end of the article
*Equal contributor

Abstract

Classification: Classification is a supervised learning method in which the computer program learns from the available data set in order to make new observations or classifications. The different classification methods used are Logistic Regression, Support Vector Machines, Random Forest and Convolutional Neural Networks.

Goal: The first goal of the project is to perform exploratory data analysis and classification on the Cleveland dataset and also calculate the performance evaluation metrics. The second goal is to train the BreakHis dataset using neural networks and identify whether the input sample belongs to benign or malignant cancer type.

Methods and Datasets: In the first project, support vector machines, logistic regression, and random forest were the 3 classifiers used to train the dataset to calculate the metrics. Whereas in the second project, Convolutional Neural Networks was used to train the dataset to achieve the final results. The classification results are presented using the publicly available Cleveland Heart Disease data set and BreakHIS data set.

Result: In the first project, we successfully achieved high accuracy in predicting the test labels and were able to plot the ROC curve for the same, however in the second project, the loss function for the training and test dataset showed contradictory behavior.

Personal key learnings: Implemented the Synthetic Minority Oversampling Technique to balance the classes.

Estimated working hours: 28 hours per week

Scientific Background

Classification is the process of identifying and grouping data into previously assigned categories [1]. In data science, classification is a process of identifying categories and placing them into a collection of data sets to perform more detailed analysis [2]. To work with the classification process, a workflow is needed. The first step is always data exploration. The second step is the pre-processing of the data and the splitting of the data into test and training data, and finally the selection of a classification method. In the Data Science block course, we studied classification problems such as the binary classification problem, Image classification, and imbalanced classification. In the first Project used the Cleveland heart disease dataset from the UCI machine learning repository. The data set has structured type (integer). Blood flow to the heart muscle is affected in coronary artery disease due to the accumulation of plaque in the arteries of the heart. It is the most common form of heart disease. Initially, invasive coronary angiography is the main reason for establishing the presence, and severity of heart disease. The goal is to develop a computerized diagnostic

tool that could replace the current invasive gold standard with the combined results of noninvasive tests and other patient features [3]. The second data set comes from research on the classification of histopathological images in breast cancer. The mortality rate of breast cancer is very high compared to other cancers, and moreover, breast cancer is one of the most common cancers in women. Detection and diagnosis of breast cancer can be done by imaging. The digitized images(unstructured data) in this research are labeled "benign" or "malignant", and the goal is to develop a tool that clinicians can use to decide whether a sample is benign or malignant [4].

Goal

In the first project, After conducting exploratory data analysis, Our model should handle missing data by clearing rows with missing values. The three different classifiers should give good accuracy, the model will tell us the difference when we consider all the labels for training vs two labels ("heart disease and no heart disease" i.e Binary Classification) for training. To analyse the performance of our model we will also plot ROC curves for each of the three classifiers. For the second assignment, we cleaned our data-set manually by adding the subcategories of benign and malignant into training and testing files, we even dismissed some images with different shapes, we will be using three different deep learning predictions model with hidden layers and one of the models is CNN.

Data and Preprocessing

The Cleveland database (1988) [3] contains 303 instances with 75 feature and 1 target attribute. It is the only We used a subset with 13 feature and 1 target attribute. The 13 features include results from non-invasive clinical tests, which are measured in interval (age, trestbps, chol, thalach, oldpeak) or nominal scale (sex, fbs, restecg, cp, exang, slope, thal, ca). The target attribute "goal" is in ordinal scale from 0 (<50% diameter narrowing) up to 1-4 (>50% diameter narrowing). In the data exploration we detected six missing values. We deleted those entries and reduced our dataset to 297 instances. The correlation between the feature attributes is low, that means they are suitable for a classification (figure 1).

We modified the target attribute "goal" into a binary outcome by changing it to whether a heart disease is present (1) or not present (0). That balanced our target attribute, the difference is visible in figure 2. Because of highly imbalanced data in some attributes we used data augmentation to balance the minority classes. Than we split the data into 30% test data and 70% training data.

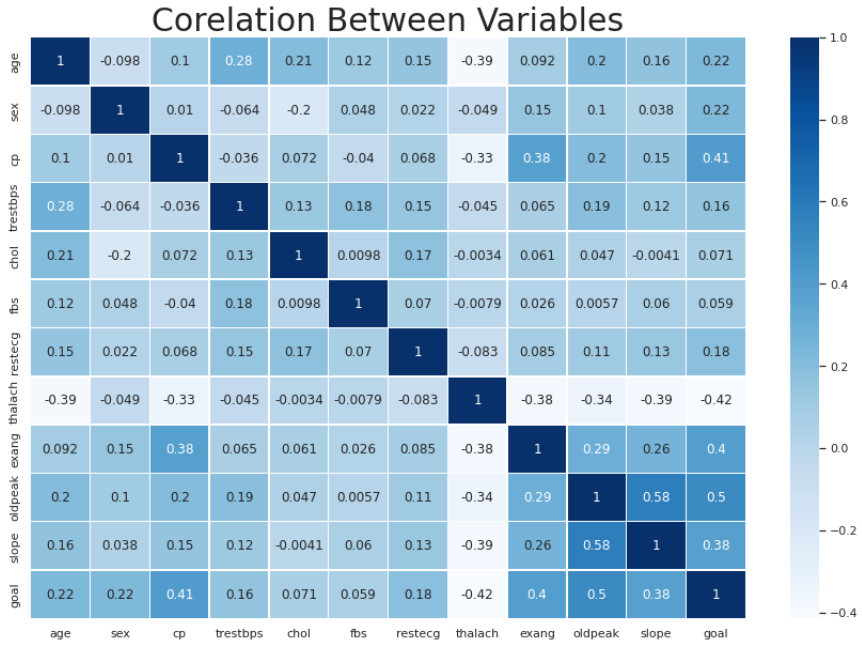

Figure 1 Correlation between feature attributes

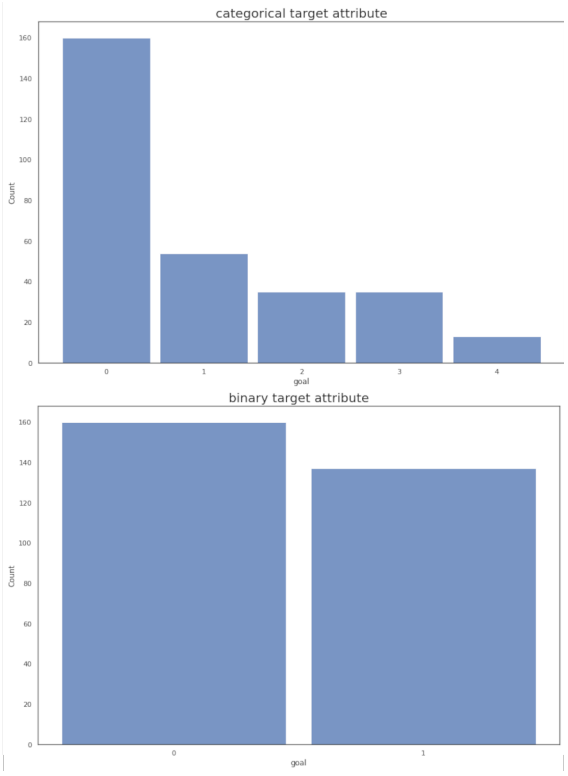

Figure 2 Difference between categorical and binary target attribute

The BreaKHis database (2015) [4] contains 7909 images of breast tumor tissues with different magnifying factors ($40x$, $100x$, $200x$, $400x$). There are four benign types of cancer: adenosis (A), phyllodes tumor (PT), fibroadenoma (F) and tubular adenoma (TA); and four malignant tumors: ductal carcinoma (DC) , lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC). We decided to train our model on a magnifying factor of $400x$. Therefore we extracted the images of that factor into a new directory. We resized the images to $200x200$ in order to make all the same resolution. Than we split the data into 25% test data and 75% training data.

Methods

In the first project we used three classifiers: support vector machines(SVM), logistic regression, and random forests. The goal was to train classifiers in the frame of a supervised machine learning project. Unfortunately, we weren't able to handle the technical difficulties in adapting the code we had access to in the task of multi-label classification of heart disease. Instead, we were led to transform the task into a binary classification one. In the case of SVM it is about finding a decision boundary to separate different classes and maximise the margins. We used a non-linear kernel (RBF) allowing for non-linear decision boundary. Logistic regression is based on the probability of an event to happen given a threshold. In the case of random forests an ensemble of decision trees is considered each of them providing a decision tree based output for classifying an instance, the ultimate goal being to assign the class according to the majority vote of all the trees.

In the second project we used convolutional neural networks (CNN) for classification of histopathological images in a multi-label context. We chose the $400x$ magnification and selected randomly 100 images from each of the 8 folders, i.e. 4 for the benign and 4 for the malignant case. We then resized the images to 200×200. By imposing different values for the number of layers, the kernel size and the rectified linear unit (relu) as activation. We chose the softmax layer for the output of the probability distribution. We also used sparse categorical crossentropy for the loss function and the accuracy as the main metric.

Results

All the classifiers used in the first project achieved high accuracy in predicting the test labels. The confusion matrices were: in the case of svm, $TP = 36, TN = 154, FP = FN = 5$, in the case of logistic regression, $TP = 32, TN = 146, FP = 13, FN = 9$, and in the case of random forests, $TP = 37, TN = 148, FP = 11, FN = 4$. The accuracy on the training and test data set being similar, we do not think that substantial overfitting would have occur. As shown in the figure 4, the ROC curves for all the classifiers demonstrate a high area under the curve as a measure of high sensitivity and specificity.

In the second project we did not achieve the goal of getting highly accurate classifiers using deep neural networks. While the loss function dropped in case of the training set it rather increases and oscillates in the case of the test set, the accuracy behaving in the opposite direction. We do not show the corresponding plot here.

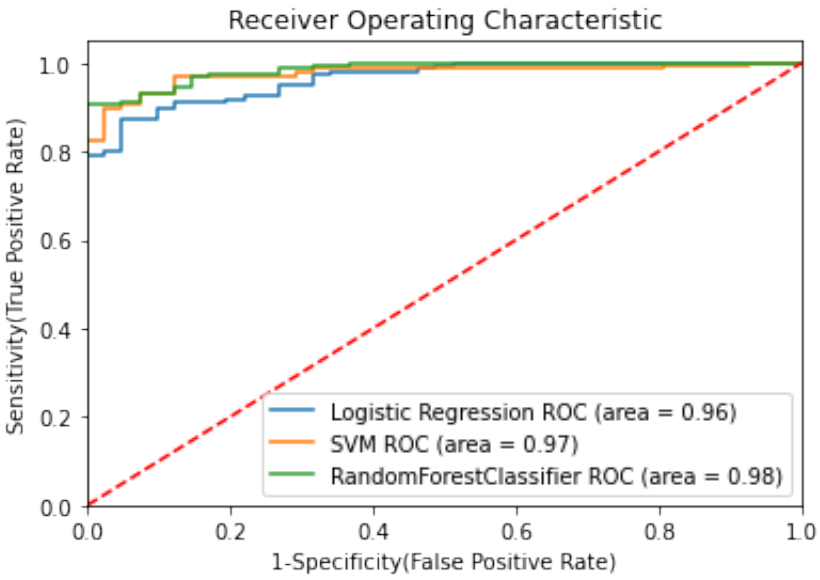

Figure 3 ROC curves of the three classifiers trained in project 1.

Discussion

Predicting a disease such as heart disease or breast cancer cannot simply be determined by only considering risk factors. Machine learning techniques such as classifications, which we have also used in these projects, can be used to predict outcomes from existing data. In the exploratory data analysis of the first project we found that certain features were stronger correlated with heart disease such as age, sex (male) and cholesterol level in blood. We handled missing values by simply discarding the corresponding instances from the data set given that only a few such instances were affected. Moreover, we used Synthetic Minority Oversampling Technique (SMOTE) to balance the classes. We trained three classifiers, SVM, logistic regression, and random forests. All of them showed high specificity and sensitivity on the test data set as shown by the corresponding confusion matrices. In particular, the corresponding ROC curves showed similar performance for the three classifiers, the area under the curve being over 96 %. In the case of SVM we observed a stronger discrepancy between the performance on the training and test data set, i.e. the prediction on the training set being higher than on the test data, suggesting that some overfitting can not be excluded. Nevertheless, more data are needed to validate the performance of the classifiers. Using such a model in predicting disease can only be made for making easier the decision making for young physicians and not for substituting them.

In the second project, we experienced different difficulties in appropriately handling the data. The goal being to train deep learning models for classification of histopathological images, we opted to use the highest magnification in order to achieve higher performance. Our goal was to train classifiers that will be able to make multi-label predictions and not binary ones. The original size of the images and the total volume of the data provided great difficulties to process them in both

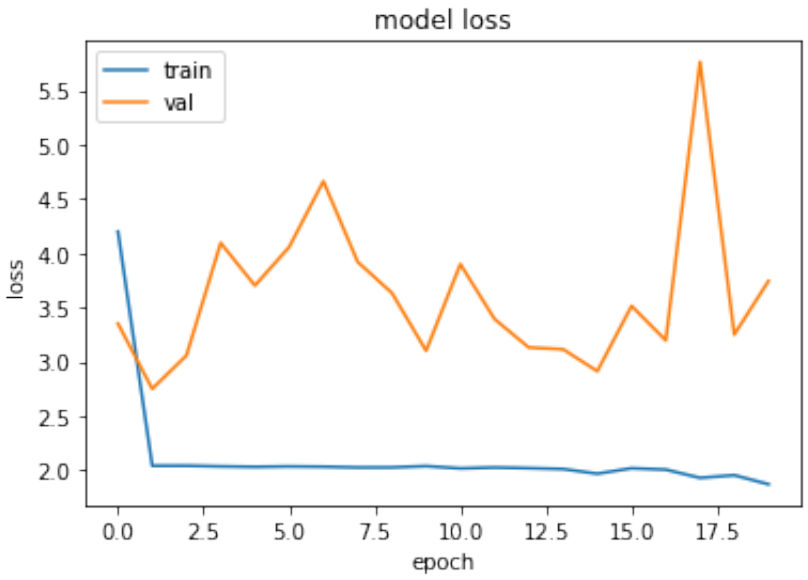

Figure 4 The loss function of one of the CNN networks.

google colab as well as in a local machine, the main issue being crashing of sessions and kernels. We thus resized the images. In addition to that we selected for the same reason only 100 images from each class. We generated three CNN based models by selected different layers and kernel sizes. With this reduction in mind, the performance dropped sensibly and the loss function increased correspondingly in all three deep learning models. We used 20 epochs for each model, but it did not sensibly change the overall performance of the models. We used code from the book by Francois Chollet and we do not think that it was due to the code rather then to the low volume of images we took into account and to the fact of resizing them to a considerable measure and thus affecting the overall prediction power.

Appendix

Shipra : Worked on Abstract and Exploratory Data Analysis.
Melika : Worked on Scientific Background and Evaluation Metrics of each model.
Dennis : Worked on Data Preprocessing and Variation of CNN model.
Neeraj : Worked on Goals of the project and Preprocessing the data sets.
Ibrahim : Worked on Results, Methods and Implementation of CNN models.

References

1. Accessed: 16.11.2022. https://www.techopedia.com/definition/13779/classification
2. Accessed: 16.11.2022. https://www.indicative.com/resource/classification-analysis/
3. Accessed: 16.11.2022. https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset
4. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering **63**(7), 1455–1462 (2015)