Classification of Pigmented Skin Lesions

Cyrus Thompson

Abstract

For many conditions such as skin cancer that can be expressed through pigmented skin lesions, a timely diagnosis can be the difference between life and death. With that in mind, in this paper we explore many possible classification models and their success in classifying a pigmented skin lesion into a class based on its underlying condition and or simply classifying if the underlying condition is life threatening . The dataset used in this paper for model training and testing is HAM10000 and the metric used to determine model success was classification error.

Introduction

The prevalent diseases such as Melanoma that can be expressed through pigmented skin lesions are among some of the leading causes of death worldwide and sadly many of these deaths could have been avoided with an earlier diagnosis. The ability to do automated diagnoses would allow for a patient to get an early diagnosis or simply a second opinion. What we are going to consider a diagnosis in the purview of this paper is the classification of features describing a pigmented skin lesion into a class which represents the condition that is causing the pigmented skin lesion. To this end the automated classification of pigmented skin lesions is an ongoing research topic with a wide range of opinions on implementation. With this in mind the trial of multiple machine learning tools would be useful for determining the potential of many of the current approaches.

Technical Approach

The main focus of this project was three approaches to the problem of classification, Logistic Regression, Support Vector Machines and Neural Nets. Logistic regression is a probability based classifier that classifies based on the likelihood the data point is in one class versus another. Support Vector Machines create a hyper plane that divides the higher level data into different classes. Neural nets classify based on values of output nodes which are obtained through use of a combination of multiple layers of neurons which are functions that compute a nonlinearity from a linear combination of input features. Within each of these methods there are a wide array of hyperparameters that are customizable which shall be covered in the next section. Beyond this the other main tool used was feature and data manipulation, as shall be covered in detail later the format data itself was not very conducive to training to begin with and thus the existing features required modification, new features had to be created and some data had to be removed.

Experimental Results

The dataset used in this paper is HAM10000. This is a fairly recently created internationally sourced dataset that was used in the ISIC 2018 classification challenge. Each patient which represents one data point has 5 pieces of associated data we utilize. The diagnosis , the age of the patient, the sex of the patient, where on the body the pigmented skin lesion occured and a image of pigmented skin lesion. The two pieces of data in the dataset that were not used for training were how diagnosis was validated and the lesion id. This is because I don't believe they are applicable for classification, but the lesion id was useful for getting rid of duplicates in the data. The diagnosis "dx" in the data which is one of seven values shall be the labels and the other pieces of data shall be the features. The possible diagnosis values are "akiec" with 327 occurences in the unfiltered dataset which represents Actinic Keratoses and Intraepithelial Carcinoma. "bcc" with 514 occurences in the unfiltered dataset which represents Basal Cell Carcinoma. "bkl" with 1099 occurences in the unfiltered dataset which represents Benign Keratosis. "df" with 115 occurences in the unfiltered dataset which represents Dermatofibroma. "nv" with 6705 occurences in the unfiltered dataset which represents Melanocytic Nevi. "mel" with 1113 occurences in the unfiltered dataset which represents Melanoma. "vasc" with 142 occurences in the unfiltered dataset which represents Vascular Skin Lesions. Yet the first thing that had to be done with the data set was to eliminate any duplicates that exist in the unfiltered dataset, creating a filtered dataset. Yet as seen above and below there was still quite a skew in the data towards "nv" Melanocytic Nevi. This led to many of the methods for classification defaulting to that class. This would even be worse in any application using these classifiers because Melanocytic Nevi are harmless and other classes such as melanoma are deadly thus a false positive in Melanocytic Nevi would be horrible. To solve this issue another dataset was created that randomly took out data in the class with the most data points till the size of that class was comparable to the smallest class.

Number of data points in each class	Unfiltered Dataset	Filtered Dataset	Manually Not Skewed Dataset
'df'	115	73	73
'bkl'	1099	718	411
'nv'	6705	5361	379
'vasc'	142	98	98
'mel'	1113	613	376
'akiec'	327	228	228
'bcc'	514	327	327

The next thing that was done to both filtered datasets was turning the values of all features and variables into numerical values. First the sex was turned into a binary value with male being 1 and all other values in the sex column being 0. Next the localization values became

their own features. Thus if the Pigmented Skin Lesion was on the neck of the patient the value for the feature neck would be 1 and the value for all other locations which are now also features would be zero. The next thing that was done is the age values were normalized to be between zero and one. Next the output class/label “dx” which represents the diagnosis or the underlying conditions were changed into numerical values such that 0='df', 1='bkl', 2='nv', 3='vasc', 4='mel', 5='akiec', 6='bcc'. This was to make identification of whether the condition was harmless easy, thus the conditions with less than 3 are not deadly and conditions of 3 or greater are deadly. The final thing that was required was to import the image information into the dataset, this was done two ways. For the Logistic Regression Support Vector Machines and Binary Neural Nets the image data was loaded in as an array of rgb values all at once. For the more advanced Neural Nets the image data was loaded dynamically. An example of the training dataset is below.

dx	age	sex	neck	lower extremity	back	trunk	abdomen	face	upper extremity	foot	genital	unknown	scalp	chest	hand	ear	acral	image_data
6	0.529412	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[[183, 137, 136], [190, 148, 153], [212, 174, ...
2	0.294118	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	[[174, 137, 147], [175, 136, 140], [176, 137, ...
2	0.588235	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	[[224, 212, 221], [225, 209, 220], [226, 210, ...
2	0.588235	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	[[226, 153, 162], [228, 153, 168], [227, 150, ...
2	0.352941	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	[[191, 112, 124], [188, 112, 130], [171, 99, 1...

Yet beyond this we hoped to achieve greater accuracy by manipulating the data above to be more easily classified. The first artificial feature added was information on the mean value and the variance of the red, blue and green colors in each image and thus creating another dataset, which had all of the above data for each patient along with the new extracted information as can be seen below. The reason for this addition was that change and differences in color underlie certain conditions and that the information of change in relation to the mean could potentially be used by the models to learn something extra.

red_var	green_var	blue_var	red_mean	green_mean	blue_mean
313.420837	824.873775	855.493453	212.304688	170.314453	174.678711
602.308722	820.310303	1083.954575	170.864258	128.390625	126.183594
100.803188	498.742942	949.542502	228.869141	195.668945	215.206055
663.335934	909.795653	1057.270019	214.498047	140.874023	144.749023
823.856430	1499.055603	1894.978848	182.621094	107.570312	116.275391

The next dataset was a dataset where the output class “dx” was changed to binary based on whether the underlying condition was life threatening. We hoped that if the models could achieve a higher accuracy on this dataset than in application it would at minimum allow a patient to know if the Pigmented Skin Lesion was life threatening and thus convenience them to see a doctor immediately and thus lower mortality of many of these underlying conditions. The final dataset was a dataset that just had the image data and was only used for the Convolutional Neural Networks. The final thing done in terms of data manipulation was a split for each of the above datasets into training, testing and validation sets. This was done by splitting the datasets which are all randomly ordered into three smaller sets. 70% of the data went to the training set 20% to the testing set and 10% to the validation set.

The first method that was used on all datasets was Logistic Regression. The main hyperparameter that needed tuning was the strength of the regularization. So the way the best Logistic Regression model was determined on any data set was 4 Logistic Regression models were trained on the training data each with a different regularization term from strong to weak. The model with the lowest classification error for the test data was then tested against the validation and the classification error for this model on the validation set, test set and training set was reported.

The second method that was used on all datasets was a Linear Support Vector Machine. The main hyperparameter that needed tuning was again the strength of the regularization. So the way the best Linear Support Vector Machine model was determined on any data set was 4 Linear Support Vector Machine models were trained on the training data each with a different regularization term from strong to weak. The model with the lowest classification error for the test data was then tested against the validation and the classification error for this model on the validation set, test set and training set was reported.

The third method that was used on all datasets was a General Support Vector Machine. The main hyperparameter that needed tuning was the kernel being used. So the way the best Linear Support Vector Machine model was determined on any data set was 3 Linear Support Vector Machine models were trained on the training data each with a different kernel including polynomials, sigmoidal and rbf. The model with the lowest classification error for the test data was then tested against the validation and the classification error for this model on the validation set, test set and training set was reported.

The fourth method which was only used on datasets with binary output classes was the Binary Fully Connected Neural Net. The main hyperparameters that needed tuning were the number of nodes per hidden layer and the number of hidden layers. So the way the best Linear Support Vector Machine model was determined on any data set was 9 Binary Fully Connected Neural Net were trained on the training data each with a different number of nodes per hidden layer or a different number of hidden layers. The model with the lowest classification error for the test data was then tested against the validation and the classification error for this model on the validation set, test set and training set was reported.

The fifth method which was only used on datasets that did not have binary output was the convolutional neural network section. This section's hyperparameters are similar to the last one in the fact that it includes the number of nodes in the hidden layers and the number of hidden layers, but also includes things such as the types of layers thus I shall call this hyperparameter in general network architecture. So for this section similar to the previous three different network architectures were trained on the training data and the one

with the best result on the testing data was tested against the validation data and the classification error for this model on the validation set, test set and training set was reported. Yet for the unskewed data the models always simply outputted “nv” thus had the same error in the training set and validation set.

Filtered Only DataSet	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Logistic Regression	18.027734976887%	26.819407008086%	29.514824797843%
Linear SVM	12.596302003081%	27.156334231805%	28.840970350404%
General SVM	0.0 %	27.291105121293%	29.514824797843%

Filtered and Color Features DataSet	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Logistic Regression	16.409861325115%	26.954177897574%	26.684636118598%
Linear SVM	20.781972265023%	36.792452830188%	35.983827493266%
General SVM	0.0 %	26.482479784366%	26.954177897574%

Filtered and Color Features and Binary DataSet	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Logistic Regression	12.268875192604 %	16.307277628032%	17.385444743935%
Linear SVM	14.271956856702%	15.566037735849%	16.442048517520%
General SVM	0.0 %	16.03773584905 %	17.250673854447%
Binary NN	12.57704160246%	14.690026954177 %	16.442048517523 %

Filtered OnlyDataSet Not Skewed	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Logistic Regression	19.729932483128%	54.593175853018%	56.544502617801%
Linear SVM	6.826706676669 %	56.430446194225%	56.020942408376%

General SVM	0.0 %	50.918635170603%	49.214659685863%

Filtered and Color Features DataSet Not Skewed	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Logistic Regression	16.654163540885%	53.280839895012%	57.06806282722513 %
Linear SVM	7.1267816954238%	54.330708661417%	56.544502617801%
General SVM	0.0 %	49.343832020997%	48.167539267015%

Filtered and Color Features and Binary DataSet Not Skewed	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Logistic Regression	19.654913728432%	33.858267716535%	36.649214659685%
Linear SVM	15.378844711177%	40.157480314962%	39.267015706806%
General SVM	0.0 %	33.333333333333%	33.507853403141%
Binary NN	26.556639159789%	30.18372703412%	38.219895287958%

Just Image Dataset	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Convolutional Neural Network	24.53987730061 %	27.291105121293%	29.514824797846%

Just Image Dataset Not Skewed	Classification Error of Returned Model On Training Data	Classification Error of Returned Model On Test Data	Classification Error of Returned Model On Validation Data
Convolutional Neural Network	76.19047619061%	76.719576719576%	81.578947368421%

Reflecting on the above results we have come to the following conclusions. The different datasets did lead to vastly different results and different methods did better on different datasets. To begin we feel the skewed data is not a good representation to see accuracy. This can be seen in the extreme in the case of the convolutional neural nets which when given the skewed data resulted in the models just returning the most common class almost always. Secondly in terms of the results of the models that used the data that was not

skewed we believe that the type of model with the highest performance was Support Vector Machines. This is because of the consistently better performance on the validation data which we believe shows that they were able to better represent the underlying data distribution. The biggest disappointment by far was the convolutional neural nets on just the image data. There are a few reasons why they might have performed so poorly. First it could be that the data other than the image such as age and location on the body are integral for classification or it might simply be a difficulty of poor network architectures compounded by computational limitations on our end. Either way convolutional neural nets by far had the worst classification accuracy which was still much better than random guessing. Logistic regression on the other hand did almost as well as the SVM and the fully connected binary neural nets but always did marginally worse. The binary fully connected neural nets on the other hand did well even comparable to SVM. On the topic of the binary classification while the models put out accuracies that were better than guessing, it was not by a large percentage wise amount. This could mean that features of the deadly and not deadly conditions are hard to distinguish; further experiments would be required. In terms of the added color features while they did lower classification error on average the amount was not significant and thus similarly this requires more exploration. This dataset is considered a challenge even to the leaders in the machine learning academia and industry and we were happy to get a chance to train with it. Yet the work and exploration does not end here, we are truly interested in further exploration. Specifically we would wish to test more models against the binary output datasets for a few reasons. The first reason is that we feel like this approach would be the most applicable for an application used by the layman, because what matters to them is identifying if the Pigmented Skin Lesions on their body is dangerous. If we were to create or inspire the creation of such a model that had a very low classification error in the binary case we really do believe that this would lower the mortality rate for some of the underlying conditions. The second reason is that no one else that we could find has approached this problem in this way and we would love to be able to report new findings. The final reason is that we really enjoyed the exploration covered in this paper even though the results could have been better.

Participants Contribution

All Code provided by - Cyrus Thompson
This Paper provided by - Cyrus Thompson
The knowledge required to do all this provided by - Ehsan Elhamifar

Thank you for a good semester stay safe and healthy

References

1 Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: “Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)”, 2018; https://arxiv.org/abs/1902.03368

2 Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).