KNN vs LR and SVM for Skin Cancer Classification

Tshifhiwa Mavhona
School of Computer Science and Applied Mathematics
University of the Witwatersrand)
Johannesburg, South Africa
1613720@students.wits.ac.za

***Abstract*—Skin Cancer Classification is the process of labeling images that contains human skin as whether it is Benign or Malignant. This means to decide if an input image contains cancerous or non cancerous cells. Traditional Machine Learning algorithms such as K-Nearest Neighbor(KNN), Logistic Regression (LR) and Support Vector Machine(SVM) can be used for this type of classification, which is a binary classification. In this paper we Investigate the performance of KNN, LR and SVM by measuring how well each model is able to predict correct labels for testing input images. We find that KNN has the least performance in accuracy of results as compared to SVM and LR with LR having the best performance of the three. Although the algorithm do not show that significant much difference performance**

***Index Terms*—KNN, SVM, Logistic Regression, Skin Cancer**

I. Introduction

Due to Increase use of radiation in our everyday life, The also have been an increase in skin cancer cases of the world. This is because Ultraviolet radiation is one of the main course of skin cancer[1]. Skin Cancer has one of the highest moralities in this world. One way in which this may be prevented is by early detection. The use of machines for clinic purposes such as cancer research, heart deceases and brain tumors (a form of cancer) has proven to be very useful over the past years[3]. With the help of machine learning we have been able to predict from a given input image whether it represent image in which the sample skin has cancer cells. Using binary classifier we can be able to produce very good results although when working with images one must take into account the dimensional of the problem, since they tend to be very high. For example if the input image is of size 28×28 then it means if we were to represent each pixel as an independent feature we would have 784 features. That's if the image is a gray scale image. For a RGB image the image would then be 3 times the value of the gray scale image creating a very large feature vector. Methods such as Principal Component Analysis (PCA) can be used to reduce the dimensions of the feature while maintain some accuracy from the original data. This actually will reduce the running time of our models. Examples of binary classification method which are can be used to solve this problem are K-Nearest Neighbor, Logistic regression, Support Vector Machine and Artificial Neural Networks.In this paper we compare the perforce of accuracy obtained by the first three algorithms. It is known that KNN tends to suffer in a data set that has a very large number of features. As discussed above we tackle this problem using PCA.

II. Methody

A. Principal Component Analysis

As discussed before PCA is a method which is used to simplify the complexity in high-dimensional data. when it transforms the data into fewer dimensions, it keeps the trends and pattern information of the data. This method is when pre-processing data. It was works by first finding the covariance matrix of the matrix which represent the data to me used. For Example if the original data produces a an $N \times M$ matrix, with M representing the number of dimensions of our data and N representing the size of our data then using PCA we find the covariance matrix using the equation

B. K-Nearest Neighbor

This algorithm works works by first taking samples of images assuming they have been pre processed and represent them as points. this points are used to learn. That is taking any new points which is not from the sample points we are to classify that point using the output of it's nearest neighbors. How we measure the nearest neighbour of a point is up to us , in our case we use the euclidean distance of feature vectors

$$D = \sqrt{\sum_{i=0}^{N} (x_i - y_i)^2}.$$

In this case D is the distance between point x and y, N is the dimensions of the feature vector and x and y denote the points which are in $\mathbb{R}^N$ dimension space. The best value of K is chosen by first running KNN on the training data for different types of K. usually the optimal value of K is between $3 - 10$ with K being an odd number

C. Logistic Regression

Unlike most regression algorithms logistic regression used to measure the probability of an input input which belongs to a binary class. It can be thought of as an extension from linear regression. There are certain conditions which must be met by the input data for logistic regression to work, with the first and obvious one being, that data must be of a binary class, that it belongs to two clusters. There also should be complete or a very high correlation between the independent variables of the feature vector.

D. Support Vector Machine

Support vector machine is an algorithm that can be used for both as regression and classification purpose, it therefore falls under the supervised learning section. In this paper we focus on the clarification part of it. The e idea around SVM is that we are looking to find a hyper-plain which best segments data points into two classes. an image illustrating this is show below:

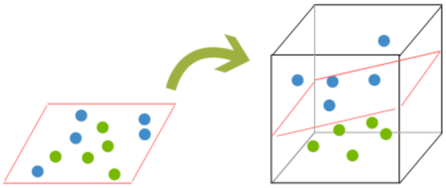

Fig. 1. Example of a hyper-plain in 3-dimensional space.

III. Related work

Much of the resent work in skin cancer detection has been focused in the use of multi tier neural networks.A feed Forward neural network is one good example in this case. firstly to be able to predict what an input image represents, there must be there use of feature extraction using image processing techniques that way it is possible to reduce the size of the feature vector to be used for classification. The KNN classifier out performs the neural network according to M. Elgamal [3]. In the the research carried by Murugan [4] it clearly outlines that the SVM method which they have used for classifying skin lesions performs better than Random forest and KNN.

IV. Experiment

V. Pre-Processing

The data comes in a form images representing benign and malignant skin condition and our aim is to create models that will be able to categorize new input images as the correct label. Examples of the two classes of our data is shown below:

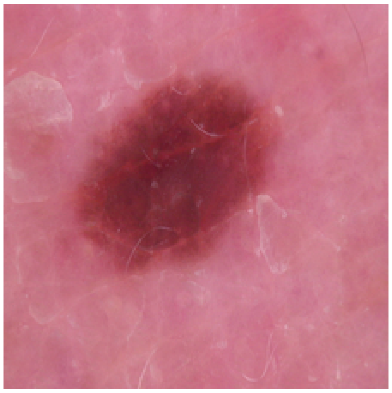

Fig. 2. Example of a benign.

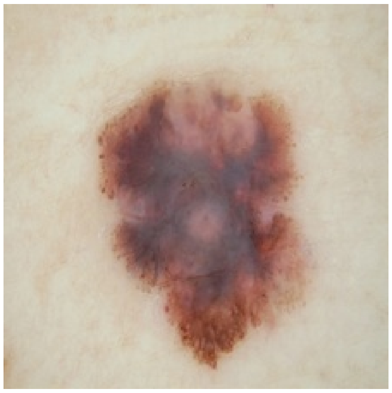

Fig. 3. Example of a malignant.

The input images start of as of size 224×224 and are reduced to a small size such we decrease the feature vector of the data. Ideally to make sure that the KNN algorithm is able to run since running a KNN on a large data set that has large features will create a lot of delay in training. The reduced data is of size 28×28 although the images are in RGB so that means the feature vector would be 3 times large. we use PCA in the second experiment to reduce the dimension of the feature vector to 20 while loosing accuracy of about less than 10%.

A. Results

Table 1 shows the confusion matrix that shows the performance of the logistic regression algorithm. The hyper-parameter which was chosen here in the number of iteration which the gradient decent uses to converge. The value of that is set to 2000.The overall performance of the logistic regression algorithm

Tbale 2 shows the confusion matrix for SVM and we can see that it didn't differ that much from that of Logistic regression. The KNN algorithm had a accuracy performance of 78.3% which is lower than that archived by both the logistic regression and SvVM since they archived an accuracy slightly higher.

TABLE I
CONFUSION MATRIX FOR LOGISTIC REGRESSION

	predicted bengin	predicted malignant
actual begin	289	71
actual malignant	62	238

TABLE II
CONFUSION MATRIX FOR SVM

	predicted bengin	predicted malignant
actual begin	292	74
actual malignant	68	226

VI. CONCLUSION

Supervised learning algorithms can be used for skin cancer detection, algorithms such as KNN, SVM and Logistic regression can be used for binary classification. after running the 3 listed algorithms we find that they do not really have that much difference in performance when used for classification. They all produce an accuracy of 78% if we were to chop from the decimal point. The use of feature extraction techniques might play a role in the performance of the classifiers this is something that may be invested in the future.

REFERENCES

[1] Narayanan, D.L., Saladi, R.N. and Fox, J.L. (2010), Review: Ultraviolet radiation and skin cancer. International Journal of Dermatology, 49: 978-986

[2] P. Yuan, Y. Chen, H. Jin and L. Huang, ”MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification,” IEEE International Workshop on Semantic Computing and Systems, Huangshan, 2008, pp. 133-140

[3] International Workshop on Semantic Computing and Systems, Huangshan, 2008, pp. 133-140 Elgamal, M., 2013. Automatic skin cancer images classification. IJACSA) International Journal of Advanced Computer Science and Applications, 4(3), pp.287-294.

[4] Murugan, A., Nair, S.H. & Kumar, K.P.S. Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers. J Med Syst 43, 269 (2019)