MACHINE LEARNING MODELS FOR THE DETECTION OF HUMAN EYE DISEASE

K. Arunkumar
Department of Computer Science, Annai Vailankanni Arts and Science College, India

Abstract

Diabetic Irish (DI) is human eye disease among people with diabetics which causes damage to irish of eye and may eventually lead to complete blindness. Detection of diabetic Irish in early stage is essential to avoid complete blindness. Effective treatments for DI are available though it requires early diagnosis and the continuous monitoring of diabetic patients. Also many physical tests like visual acuity test, pupil dilation, and optical coherence tomography can be used to detect diabetic Irish but are time consuming. The objective of the study is to give decision about the presence of diabetic Irish by applying ensemble of machine learning classifying algorithms on features extracted from output of different irishl image. It will give us accuracy of which algorithm will be suitable and more accurate for prediction of the disease. Decision making for predicting the presence of diabetic Irish is performed using Random Forest, K-Nearest Neighbor, Neural Networks and Support Vector Machine.

Keywords:

Random Forest, K-Nearest Neighbor, Neural Networks and Support Vector Machine

1. INTRODUCTION

. Diabetes is a chronic and organ disease that occurs when the pancreas does not secrete enough insulin or the body is unable to process it properly. Over time, diabetes affects the circular system, including that of the irish. Diabetes Irish (DI) is a medical condition where the irish is damaged because of fluid leaks from blood vessels into the irish. It is one of the most common diabetic eye diseases and a leading cause of blindness. Nearly 415 million diabetic patients are at risk of having blindness because of diabetics. It occurs when diabetes damages the tiny blood vessels inside the irish, the light sensitive tissue at the back of the eye. This tiny blood vessel will leak blood and fluid on the irish forms features such as micro-aneurysms, haemorrhages, hard exudates, cotton wool spots or venous loops. Diabetic Irish can be classified as non-proliferative diabetic Irish (NPDR) and proliferative diabetic Irish (PDR). Depending on the presence of features on the irish, the stages of DR can be identified. In the NPDR stage, the disease can advance from mild, moderate to severe stage with various levels of features except less growth of new blood vessels. PDR is the advanced stage where the fluids sent by the irish for nourishment trigger the growth of new blood vessels. They grow along the irish and over the surface of the clear, vitreous gel that fills the inside of the eye. If they leak blood, severe vision loss and even blindness can result.

Currently, detecting DI is a time-consuming and manual process that requires a trained clinician to examine and evaluate digital colthe fundus photographs of the irish. By the time human readers submit their reviews, often a day or two later, the delayed results lead to lost follow up, miscommunication, and delayed treatment.

This paper mainly focuses on the prediction of diabetic Irish and analysis performed of different algorithm for the prediction. Machine learning algorithms such as KNN, RF, SVM, NNET etc. can be trained by providing training datasets to them and then these algorithms can predict the data by comparing the provided data with the training datasets. The objective is to train the algorithm by providing training datasets to it and the goal is to detect diabetic Irish using different types of classification algorithms.

2. BACKGROUND

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data [4]. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Tom M. Mitchell provided a widely quoted and more formal definition:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [5].

The core of machine learning deals with representation and generalization. Representing the data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the ability of a machine learning system to perform accurately on new, unseen data instances after having experienced a learning data instance. The training examples come from some generally unknown probability distribution and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases. The performance of generalization is 11 usually evaluated with respect to the ability to reproduce known knowledge from newer examples. There are different types of machine learning, but the two main ones are:

- Supervised Learning
- Unsupervised Learning

2.1 SUPERVISED LEARNING MODEL

Supervised learning is the machine learning task of inferring a function from supervised training data [6]. Training data for supervised learning includes a set of examples with paired input subjects and desired output. A supervised learning algorithm analyses the training data and produces an inferred function, which is called classifier or a regression function. The function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a reasonable way.

A simple analogy to supervised learning is the relationship between a student and a teacher. Initially the teacher teaches the student about a particular topic. Teaching the student the concepts

of the topic and then giving answers to many questions regarding the topic. Then the teacher sets an exam paper for the student to take, where the student answers newer questions.

The Fig.1 describes that the system learns from the data provided which contains the features and the output as well. After it has done learning, newer data is provided without outputs, and the system generates the output using the knowledge it gained from the data on which it trained. Here is how supervised learning model works.

Fig.1. Workflow of supervised learning model

3. ALGORITHMS

Since there are so many algorithms for machine learning, it is not possible to use all of them for analysis. For this research paper, the study will be using four of them neural networks (NNET), random forest (RF), K-Nearest Neighbor (KNN) and support vector machine (SVM).

3.1 NEURAL NETWORKS

Within the field of machine learning n neural networks are a subset of algorithms built around a model of artificial neurons spread across three or more layers [10]. There are plenty of other machine learning model which is notable for being adaptive in nature. Every node of neural network has their own sphere of knowledge about rules and functionalities to develop it-self through experiences learned from previous techniques that don't rely on neural networks. Neural networks are well-suited to identifying non-linear patterns, as in patterns where there isn't a direct, one-to-one relationship between the input and output [8]. This is a learning training. Neural networks are characterize by containing adaptive weights along paths between neurons that can be tuned by a learning algorithm that learns from observed data in order to improve model. One must choose an appropriate cost function. The cost function is what is used to learn the optimal solution to the problem being solved [7]. In a nutshell, it can adjust itself to the changing environment as it learns from initial training and subsequent runs provide more information about the world.

3.2 RANDOM FOREST

Random forest algorithm can use both for classification and the regression kind of problems. It is supervised classification algorithm which creates the forest with a number of tress [9]. In general, the more trees in the forest the more robust the forest looks like. It could be also said that the higher the number of trees in the forest gives the high accuracy results. There are many advantages of random forest algorithms. The classifier can handle the missing values. It can also model the random forest classifier for categorical values [2]. The over fitting problem will never come when the study use the random forest algorithm in any classification problem. Most importantly it can be used for feature engineering which means identifying the most important feature out of the available feature from the training dataset [4].

3.3 K-NEAREST NEIGHBORS

K-nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure [11]. KNN has been used in statistical estimation and pattern recognition. KNN makes prediction for a new instance (x) by searching through the entire training set for the k most similar instances and summarizing the output variable for those k instances. For regression this might be the mean output variable, in classification this might be the mode class determine which of the k instances in the training dataset are most similar to new input many distance measure is used like Euclidean distance, Manhattan distance, Minkowski distance.

3.4 SUPPORT VECTOR MACHINE

A more formal definition is that a support vector machine constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [1].

SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user to apply a classier to data that have no obvious fixed-dimensional vector space representation [4].

4. PROPOSED MODEL FOR PREDICTION

This chapter contains proposed model, dataset collection, description, data visualization and also classifying algorithms that are used for analysis performance.

4.1 PROPOSED MODEL

The First phase is data collection. The study have collected the dataset from UCI Machine Learning repository website. The dataset contains features extracted from Messidor image set to predict whether an image have signs of diabetic Irish or not. Then features and labels of the dataset are identified. After that the dataset is divided into two sets, one for training where most of the data is used and the other one is testing. In training set four different classification algorithms has been fitted for the analysis performance of the model. The algorithms the study used are k-Nearest Neighbor, random forest, support vector machine and neural networks. After the system has done learning from training datasets, newer data is provided without outputs. The final model generates the output using the knowledge it gained from the data on which it was trained. In final phase the study get the accuracy of each algorithm and get to know which particular algorithm will give us more accurate results for the prediction of diabetic Irish.

5. IMPLEMENTATION

5.1 DATA COLLECTION

In the project the study have used a dataset that is obtained from the UCI Machine Learning Repository. This dataset contains features extracted from Messidor image set to predict whether an image contains signs of diabetic Irish or not. All features represent either a detectedv lesion, a descriptive feature of an anatomical part or an image-level descriptor. The Messidor database has been established to facilitate studies on computer-assisted diagnoses of diabetic Irish. The study have seen different kind of datasets in kaggle, github and other websites which was used for different kind of projects based on diabetic Irish. As the study wanted to work with detection of diabetic Irish, this dataset will be appropriate for the work as it has different types of features.

5.2 DATA DESCRIPTION

The dataset contains different types of features that is extracted from the Messidor image set. This dataset is used to predict whether an image contains signs of diabetic Irish or not. The value here represents different point of irish of diabetic patients. First 19 columns in the dataset are independent variables or input column and last column is dependent variables or output column. Outputs are represented by binary numbers. “1” means the patient has diabetic Irish and “0” means absence of the disease. The study have also calculated count, mean, max, standard deviation of the values in the dataset.

5.3 DATA VISUALIZATION

Another important feature in the data distribution is the skewness of each class. Data visualization helps to see how the data looks like and also what kind of data correlation the study have. The dataset distribution of each feature is shown below in figure 3.5. This is a histogram. A histogram is an accurate graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. Histograms are a great way to get to know ythe data. They allow you to easily see where a large and a little amount of the data can be found. In short, the histogram consists of an x-axis and a y-axis, where the y-axis shows how frequently the values on the x-axis occur in the data.

- As the given input variables are numeric, the study can also create box plot.
- A Boxplot typically provides the median, 25^{th} and 75^{th} percentile, min/max that is not an outlier and explicitly separates the points that are considered outliers.

5.4 SPLIT DATASET

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when separating a data set into two parts, most of the data is used for training, and a smaller portion of the data is used for testing. The study have also split the dataset into two sets. One is for training and another for testing. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. After the model has been processed by using the training set, the study have tested the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that the study want to predict, it is easy to determine whether the model's guesses are correct or not. In addition, the study have used 80% of the data for training and 20% for testing.

5.5 MACHINE LEARNING ALGORITHM

The study went through a process of trial and error to settle on a short list of algorithms that provides better result as the study are working on classification of diabetic Irish, the study used some machine learning classification algorithms. The study get an idea from the data visualizations plots which algorithms will be suitable for the classification problem. The Machine Learning system uses the training data to train models to see patterns, and uses the test data to evaluate the predictive quality of the trained model. Machine learning system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics.

So, for the study the study will evaluate four different machine learning algorithms

- Neural Networks (NNET)
- Random Forest
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)

5.6 K-FOLD CROSS VALIDATION

K-Fold Cross Validation is common types of cross validation that is widely used in machine learning. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. In the project the study used 10-fold cross validation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

6. EXPERIMENTAL RESULTS AND ANALYSIS

In the previous chapter the study have discussed about proposed system and implementation of the study. The study have demonstrated how the study collected the dataset, dataset description, visualization and algorithms the study used. Now the study discussing about the results the study obtained from the experiments upon the implementation of this system. The study have divided the dataset into two parts: training and testing dataset. In this chapter the study will show the outcome of the training and testing dataset. As mentioned before the study have used four machine learning algorithms. First, the study trained the dataset with these four algorithms and then the study built a model. Then, the study tested the testing dataset in this model. If the test set accuracy is near to train set accuracy then the study can conclude that the study built a good model.

The study have total 1151 data of different individual in the dataset. There are 1151 rows and 20 columns in the dataset. After splitting the data into two parts now the study have 920 rows for train data and for test data the study have 231 rows. When the study trained the train data for analysis performance of different algorithms. This is the result the study got-

6.1 COMPARISON BETWEEN ALGORITHMS

A comparison between the algorithms the study used for the training dataset. Here, the tall line indicates standard deviation and the rectangular box indicates median value and the brown line in the box indicates the mean value. From here the study can understand which algorithm is good for the model.

Fig.5. Comparison between algorithms

After training the model the study test the model with the testing dataset. The study have 20% data for testing in the testing set. Table.1shows the testing accuracy, precision, recall and F1 score. The detailed information of the test data evaluation with unigram model is as follows-

Table.1. Accuracy of test dataset

Models	Accuracy	Precision	Recall	F1 Score
SVM	57.07 %	62%	57%	53%
KNN	64.50 %	65%	65%	65%
RF	63.63 %	64%	64%	64%
NNET	75.32 %	78%	75%	75%

The history object is returned from calls to the fit function used to train the model. Metrics are stored in a dictionary in the history member of the object returned.

7. CONCLUSION

In experimental result, the study observe that the accuracy of the both training and testing set is quite similar and for both training and testing dataset NNET algorithm is giving higher accuracy rate which is around 75%. So, the study can say that this algorithm will give us more accurate prediction about the disease. As the main purpose of the study is to build a model which will classify the diabetic Irish as accurate as possible, the study hope that this final model will give us proper and appropriate results.

The study have also determined the train and test model accuracy and loss. For this visualization model the study have used keras package for obtaining this train and test -loss and accuracy. The study have also used History callback for this purpose. One of the default callbacks that are registered when training all deep learning models is the History callback. It records training metrics for each epoch. This includes the loss and the accuracy (for classification problems) as well as the loss and accuracy for the test dataset, if one is set.

REFERENCES

[1] T. Zheng, W. Xie and L. Xu, L., "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records", *International Journal of Medical Informatics*, Vol. 97, pp. 120-127, 2017.

[2] K. Plis, R. Bunescu and C. Marling, "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-12, 2014.

[3] B. Patra, L. Jena and S. Bhutia, "Evolutionary Hybrid Feature Selection for Cancer Diagnosis", *Proceedings of International Conference on Intelligent and Cloud Computing*, pp. 253-263, 2021.

[4] S.A. Kaveeshwar and J. Cornwall, "The Current State of Diabetes Mellitus in India", *Australasian Medical Journal*, Vol. 7, No. 1, pp. 45-48, 2014.

[5] L. Jena, S. Nayak and R. Swain, "Chronic Disease Risk (CDR) Prediction in Biomedical Data using Machine Learning Approach", *Advances in Intelligent Computing and Communication*, Vol. 109, pp. 1-13, 2021.

[6] Maryjo M George and S. Kalaivani, "Intensity Inhomogeneity Correction and Tissue Segmentation of MR Images: A Parametric Approach", *International Journal of Pure and Applied Mathematics*, Vol. 115, No. 9, pp. 409-416, 2017.

[7] Nida M. Zaitoun and Musbah J. Aqel, "Survey on Image Segmentation Techniques", *Proceedings of International Conference on Communication, Management and Information Technology*, pp. 797-806, 2015.

[8] Hui Liua, Shanshan Liu, Dongmei Guo, Yuanjie Zheng, Pinpin Tanga and Guo Dan, "Original Intensity Preserved Inhomogeneity Correction and Segmentation for Liver Magnetic Resonance Imaging", *Biomedical Signal Processing and Control*, Vol. 47, No. 1, pp. 231-239, 2019.

[9] Xiao-Feng Wang, De-Shuang Huang and HuanXu, "An Efficient Local Chan-Vese Model for Image Segmentation", *Pattern Recognition*, Vol. 43, No. 3, pp. 603-618, 2010.

[10] A. Soudi, G. Neumann and A. Van den Bosch, "Arabic Computational Morphology: Knowledge-Based and Empirical Methods", *Proceedings of International Arab Conference on Arabic Computational Morphology*, pp. 3-14, 2007.

[11] J.H. Yousif, "Natural Language Processing based Soft Computing Techniques", *International Journal of Computer Applications*, Vol. 77, No. 8, pp. 1-7, 2013.