International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 4 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1953173

Machine Learning Approach for Classification of Cancer Stages

Shubham Hingmire
School of Electronics and Communication Engineering, MIT World Peace University Kothrud, Pune, Maharashtra, India

ABSTRACT

The simplest form of health care is diagnosis and prevention. of disease. Machine learning (ML) methods help achieve this goal. This project aims to compare method of computer aided medical diagnoses. The first of these methods is a classify disease diagnosis according to their data. This involves the training of an Artificial Neural Network to respond to several patient parameters. And also comparing various classification methods the purpose research classifier classifies the patients in two class first is malignant and second is benign.

Keywords : CSV Data File Codes; Machine Learning Cancer Disease; Info Gain, Malignant, Benign

I. INTRODUCTION

Any classification method uses a set of features or parameters to characterize each object, where these features should be relevant to the task at hand. We here we consider the method of controlled classification, that is, the human-expert has determined the class that can assign the object, and also provides a set of samples of the object with the known class. This set of known objects is called a training set because the categorizer uses it to learn how to classify objects. There are two stages to building a classifier. In the training phase, training set is used to determine how to weigh and combine parameters to separate objects of different categories[3] .In the application phase the weights defined in the training set are applied to a set of objects with no known classes to determine what their class might be. ML techniques were introduced in the recent years. These Method can be used for finding frequent patterns in a huge database and derive useful knowledge from it. Classification, Association and Clustering are the Three categories algorithm in machine learning Each of these has a unique set of guidelines when applied on clinical data [4]. Effective use of these techniques will help derive significant knowledge.

II. METHODS AND MATERIAL

Classifiers

A. Neural Network

There are a number of standard classification methods in use. Probably neural network methods are most widely known. Applied neural network method to solve the problem of galaxy classification on digital photo plates range. The biggest advantage of neural network methods is that they are common: they can handle problem with a large number of parameters and can classify objects well, Even the distribution of objects in the N-dimensional parameter space is very difficult. The disadvantage of neural networks is that it is very good., especially during the training and application phases. other significant disadvantage of neural network is that it is very difficult to determine how the net is making its decision. Consequently, it is hard to determine which of the image features being used are important and useful for classification and which are worthless. As I discuss below the choice of the best features is an important part of developing a good classifier, and neural nets do not give much help in this process.

B. Logistic Regression

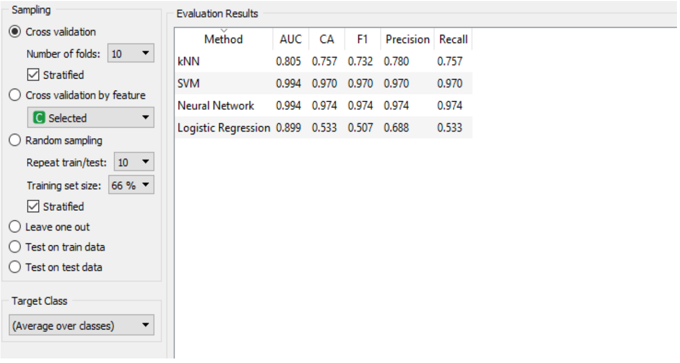
Logistic regression is a powerful statistical method that can model binomial results with one or more explanatory values (value 0 or 1, for example, the presence or absence of disease). See two main advantages of logistic regression over Chi2 or Fischer's exact test. The first is you can include more than one explanatory variable (dependent variable) and those can either be dichotomous, ordinal, or continuous. The second is that logistic regression provides a quantified value for the strength of the association adjusting for other variables (removes confounding effects). The exponential of coefficients correspond to odd ratios for the given factor.

C. Support Vector Machines

Support vector machines are conceptual planes that define decision boundaries. A decision plan is a plan that separates a set of objects with different class attributes. (SVM) is first and foremost a classification method that performs classification tasks by constructing hyperplanes in multidimensional space. Cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

D. K Nearest Neighbor

A very simple classifier can be based on a nearest-neighbor approach. In this method, one simply finds in the N-dimensional feature space the closest object from the training set to an object being classified. Since the neighbor is nearby, it is likely to be similar to the object being classified and so is likely to be the same class as that object. The advantage of the nearest neighbor method is that they are easy to implement. If you carefully select objects and carefully weigh them when calculating distances, they can also give fairly good results.. The nearest neighbor approach has several serious drawbacks. First, they (such as neural networks) do not simplify the distribution of objects in the parameter space with a clear set of parameters. Instead, the training set is completely saved as a description of the object distribution. of objects in parameter space to a comprehensible set of parameters. Instead, the training set is retained in its entirety as a description of the object distribution. There Some refinement techniques can be used in the training set, but the results are usually not a compact description of the object distribution. If there are many examples in the training set, the method is also quite slow. The most serious shortcoming of the nearest neighbor method is that they are very sensitive to the existence of unrelated parameter. Adding single parameter with random values for all objects (so it won't separate classes) will cause these methods to fail.



Sampling
● Cross validation
Number of folds: 10
☑ Stratified
○ Cross validation by feature
Selected
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Target Class
(Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.805	0.757	0.732	0.780	0.757
SVM	0.994	0.970	0.970	0.970	0.970
Neural Network	0.994	0.974	0.974	0.974	0.974
Logistic Regression	0.899	0.533	0.507	0.688	0.533

DATA DESCRIPTION

A. Data set

In order to understand more about the Medicare data, the CSV files provides with a different parameter which describes the data. Deeper understanding on this using following parameter

B. Parameters

- Area mean, Texture se
- Smoothness mean, Smoothness se

- Compactness mean, Compactness se
- Concavity mean, Symmetry e
- Concave points mean, Fractal dimension se
- Symmetry mean, Area worst
- Fractal dimension mean, Perimeter worst
- Radius se, Perimeter se.

III. Proposed Approach

The schematic diagram in results section gives the detailed design of the proposed approach. The proposed approach has two main phases training and testing. The training phase has the following steps:

A. Classification-Training Phase

Classification is a ML technique that is used to accurately predict a target class. Classification has a training phase in which the classification algorithm find the connection between the attributes and predicts the target class. There are several classification algorithms likes SVM, logistic regression, neural networks etc. This model is tested in the Testing phase

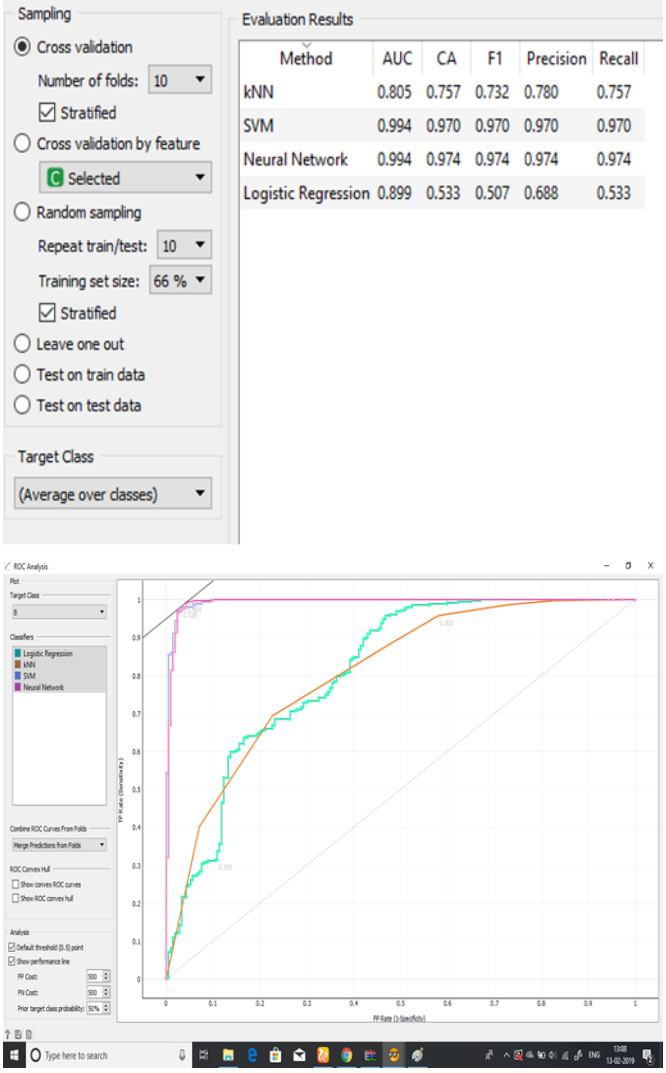
B. Testing Phase

Data from a new sample is tested. Then same steps of data definition and data extraction are followed. This data sample is then tested using the prediction model based on two performance measures which is accuracy and confusion matrix. This procedure is repeated.

IV. RESULT ANALYSIS

Result

After applying the procedure based on the ML techniques for the 11 chronic diseases, the accuracy of the classifier For training and testing phase results are different for various classifiers. For the testing data it is observed that for most of the accuracy is between 80-90%. Two stages malignant and benign have accuracy above 90% for the test data in Neural Networks.



Sampling
- ● Cross validation
  - Number of folds: 10
  - ☑ Stratified
- ○ Cross validation by feature
  - Selected
- ○ Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %
  - ☑ Stratified
- ○ Leave one out
- ○ Test on train data
- ○ Test on test data

Target Class: (Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.805	0.757	0.732	0.780	0.757
SVM	0.994	0.970	0.970	0.970	0.970
Neural Network	0.994	0.974	0.974	0.974	0.974
Logistic Regression	0.899	0.533	0.507	0.688	0.533

ROC Analysis

Plot: Target Class: B

Classifiers: Logistic Regression, kNN, SVM, Neural Network

Combine ROC Curves From Folds: Merge Predictions from Folds

ROC Convex Hull: □ Show convex ROC curves □ Show ROC convex hull

Analysis: ☑ Default threshold (0.5) point ☑ Show performance line

FP Cost: 500

FN Cost: 500

Prior target class probability: 50%

Fig 1. Training and testing data statistic

Table 1 : Result

Sr. No.	Predicted Classes		
	Total patients	**Benign**	**Malignant**
1.	569	357	212

V. FUTURE WORK AND CONCLUSION

The paper has explored the CSV dataset based on the excel file, The data was restructured and extracted based on classes diseases diagnosis codes. Then need to apply various classification algorithms to get best classification rate for each classification method was justified by analyzing the number of diagnostic tests actually done by the classifier. Then ML techniques were applied to derive the classes between malignant and benign. set of diagnostic codes were obtained for classes in the training phase. These codes were tested on a test data sample using accuracy and confusion matrix as the performance metric. It was understood that these results set of classifier gives more accurate result. The prediction model explores all data sets parameters. This research work provides the classified patients in two classes which are malignant and benign. also there possible occurrence.

VI. REFERENCES

[1]. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L, Disease Prediction by Machin Learning Over Big Data from Healthcare Communities, IEEE Access, 5, 88698879, 2017J.

[2]. Nithya, B., & Ilango, V. (2017), Predictive analytics in health care using machine learning tools and techniques, IEEE International Conference on Intelligent Computing and Control Systems (ICICCS) 2017.

[3]. Ilhan, H. O., & Celik, E. (2016), The mesothelioma disease diagnosis with artificial intelligence methods,2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT).

[4]. Shekhar, M., Chikka, V. R., Thomas, L., Mandhan, S., & Karlapalem, K. (2015), Identifying Medical Terms Related to Specific Diseases, IEEE International Conference on Data Mining Workshop (ICDMW). 2015Machine Learning Approach for Classification Of Cancer Stages

[5]. Aftab, S., Abbas, W., Bilal, M. M., Hussain, T., Shoaib, M., &Mehmood, S. H. (2013), Data mining in insurance claims (DMICS) two-way mining for extreme values, Eighth International Conferenceon Digital Information Management (ICDIM). 2013.

[6]. Kenyon, D., & Eloff, J. H. (2017)., Big data science for predicting insurance claims fraud,2017 Information Security for South Africa(ISSA) .

[7]. H. Bhavsar and A. Ganatra, A Comparative Study of Training Algorithms for Supervised Machine Learning, International Journal of Soft Computing and Engineering (IJSCE), vol. 2, pp. 74-81, 2012

[8]. Xie, Y., Schreier, G., Chang, D. C. W., Neubauer, S., Liu, Y.,Redmond, S. J., & Lovell, N. H. (2015), Predicting Days in Hospital Using Health Insurance Claims, IEEE Journal of Biomedical and Health Informatics, 19(4), 12241233

[9]. E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J.Bost, J. Tejedor-Sojo, and J. Sun, Multi-layer representation learning for medical concepts, International Conference on Knowledge Discovery and Data Mining, ser. KDD 16. New York, NY, USA:ACM, 2016, pp. 14951504.

Cite this article as :

Shubham Hingmire, "Machine Learning Approach for Classification of Cancer Stages", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 01-04, July-August 2019. Available at doi : https://doi.org/10.32628/CSEIT1953173
Journal URL : http://ijsrcseit.com/CSEIT1953173