IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Classification Analysis of Genetic Variations In Cancer Diagnosis By Multiclass Classifiers

[1]Naresh,Kumar J, [2]Sean Benhur. J, [3]Dr.S.Kanchana
[1]Student, Department of Statistics, [2]Student, Department of Software systems,
[3]Assistant Professor, Department of Software Systems,
PSG College of Arts & Science, Coimbatore-641014

Abstract: **M**achine learning in medical imaging plays the greatest disruptive technology in decades. It is being emerged not only to identify cancerous tumors at an earlier stage, but also detect and classify lesions, analyze data, reconstruct images, and more. Cancer is one of the heterogeneous disease consists of many different subclasses. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research for facilitating the subsequent clinical management of patients. Researchers, healthcare organizations, companies from biomedical and bioinformatics look forward for improving clinical outcomes for cancer patients and those who may not by using diagnostic and prognostic biomarkers. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers). Currently this interpretation of genetic mutations is being done manually based on evidence from text-based clinical literature. This article is focusing on analyzing the performance of multiclass classification algorithms for genetic features.

Index Terms - Classification, Cancer Diagnosis, Machine Learning, SGD Classifier, Support Vector Machine, Logistic regression, K Nearest Neighbour.

I. Introduction

Having well equipped science and technology still there are many places where human observation couldn't achieve a good state of art. Especially in the area of medicine where diseases caused by pathogens gets mutated at every approximate constant time. In such cases Artificial Intelligence might give us a helping hand to sort out these problems. Recent Development done by Google collaborated with Aravind eye hospital succeeded to deploy Artificial Intelligence in detecting eye disease by machines which is now considered a milestone success in the field of medicine.

It would be possible for Artificial Intelligence to successfully detect each disease in upcoming years. The most required things for successful prediction of the machine is the data. Data like clinical evidences, mutations in pathogens and more such data can make the model to predict with almost 100% accuracy. Genetic mutation of cancer with clinical evidences have been classified using the models. Varies models like Naïve Bayes, K Nearest Neighbors, Logistic Regression, Linear Support Vector Machine, Random Forest, Voting Classifier and Deep Neural Networks are used. Each model gave different results with not a lot difference in them but some models gave the least log loss compared to others. Our main goal is to reduce the log loss of the predictive model which in turn gives us better accuracy. The rest of the paper is arranged as follows: Section II reviews existing work on multiclass classification. Section III discusses the importance of preprocessing using SGD Classifiers. In section IV, state of art multiclass classification algorithms are illustrated, followed by experimental results are discussed in section V. Conclusion of the observation is specified in the Section VI.

II. Literature survey

Incredible developments in high-performance computer capabilities, machine learning algorithms can now achieve reasonable success in predicting risk in certain cancers by assessing multidimensional clinical and biological data. Machine learning approaches are enabling us to obtain explanations for patient-specific predictions. With the algorithms like Support Vector Machine, Decision Tree, Naïve Bayes, K Nearest Neighbors, Logistic Regression, Random Forest, Voting Classifier and some developed models like Deep Neural Networks, Long Short Term Memory Networks, it is possible to equalize machine to a man's brain.

QingLiao, YeDing Zoe L.Jiang proposed a novel multi-task deep learning (MTDL) method to solve the data insufficiency problem [1]. Since MTDL leverages the knowledge among the expression data of multiple cancers to learn a more stable representation for rare cancers, it can boost cancer diagnosis performance even if their expression data are inadequate.

Many researchers have emphasized the importance of AI and deep learning in healthcare for the delivery of improved quality and safety of care [4–12]. Bejnordi et al. [8] used deep learning algorithm to diagnose breast cancer tumors and compared performance with pathologists' diagnoses. Results showed that automatic detection using deep learning algorithm outperformed human diagnosis.

Haifeng et al. [13] proposed a novel method for breast cancer prediction using data mining techniques. They formulated an effective way to predict breast cancer based on patients' clinical records. They evaluated the method on two popular publicly available datasets: the Wisconsin Breast Cancer Database (WBCD) and the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. They evaluated the performance of support vector machines (SVMs), eight learning models, artificial neural networks (ANNs), Naïve Bayes classification and AdaBoost Tree. They proposed a combined model based on principal component analysis and other data mining models for feature reduction and suggested that other models such as k-mean could be used for feature space reduction.

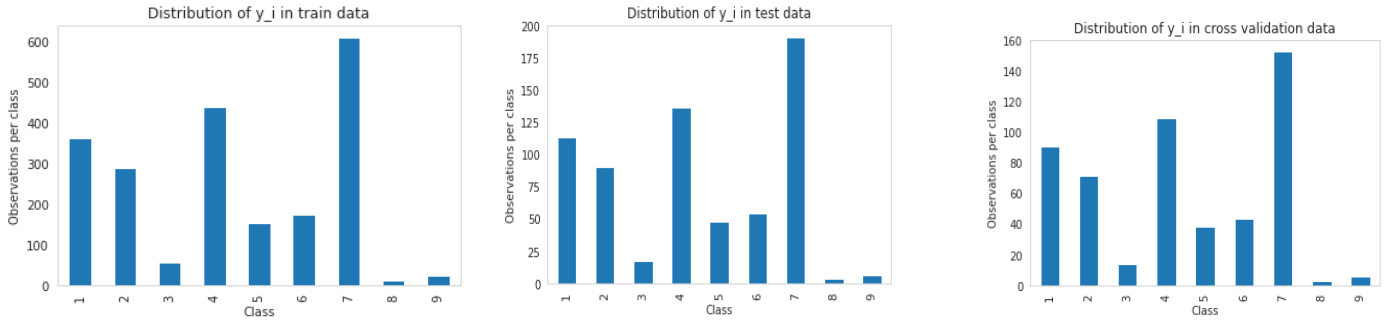
Ahmed et al. [14] used decision trees (DTs), ANNs and SVMs. Mandal et al. [15] used LR, Naïve Bayes (NB) and DTs. They also analyzed the time complexity of each classifier. LR outperforms other classifiers with the highest accuracy. However, Borges et al. [16] evaluated the performance of Bayesian networks and DT and found that Bayesian networks performed better, with 97.80% accuracy.

III. Data Preprocessing

Data preprocessing is essential for improving the performance of the classification system and will be well suitable for reducing Log Loss. *Memorial Sloan Kettering (MSK)* dataset is used for analyzing the performance of multiclass classifiers. MSK dataset consists of 3321 rows and 5 columns. Details of the data includes the id of the row used to link the mutation to the clinical evidence, the gene where this genetic mutation is located, aminoacid variation for this mutations, genetic mutation classes, and text description. The data in the text column which is the clinical trail is the main focus point as it defines a lot about the class. There are numerous method and ways to preprocess the data like One Hot Encoding, Embedding and Tokenizing. In order to make the model predictive the correct choice of data preprocessing is required to get the minimum loss.

Before getting to in depth of preprocessing of text to numeric values it is essential to preprocess with the missing data. In dataset, id with values 1109, 1277, 1407, 1639 are not having text description. These are the missing text data which are scored off from the dataset. This is done because we could not just fill the text with another existing text corresponding to same classes. Doing it leads to misinterpretation. Also since we have only few missing data we could better ignore those and move on instead of making a mess.

The next move is to check the distribution of data to each class so that when we are doing sample split in data to train part and test part we could do it with stratified random sampling method in order to avoid bias. Now being aware of the distribution of data in each class, we could stratify the data into training data, Validation data and test data. In Figure 1.1, distribution of data is clearly represented as graphs.



(a) (b) (c)

Figure 1.1 Gene distribution data

(a) Training Data (b) Test Data (c) Cross Validation Data

3.1 SGD Classifier for Preprocessing

Stochastic Gradient Descent (SGD) is an incremental, iterative method for optimizing differentiable objective function. SGD classifier selects random samples instead of a single group. SGD classifier is used for minimizing the log loss functions based in function gradient and for knowing the contribution of Gene column, Variation column and Text column of the data. In reduction of loss, Gene column of the data is subjected to fit to the SGD Classifier model and loss reduction will be observed in the model. Variation and text expression of the data is also subjected to fit to the SGD Classifier model to observe the contribution to the loss reduction in the model. The model is also subjected to tuning of hyper parameter alpha.

IV. Classification Techniques

Classification is a classic machine learning application for categorizing data and forming groups based on the similarities. Machine learning algorithms are used to determine the output of this problem, which will be either Yes or No (Two classes). Binary Classification basically categorizes output in two classes. To classify something that has more than 2 categories and isn't as simple as a yes/no problem. Multiclass classification can be defined as the classifying instances into one of three or more classes.

In this paper, state of art classification algorithms multinomial Naïve Bayes, Support Vector Machine, logistic regression and K Nearest Neighbor are used for multiclass classification of text description data in cancer diagnosis.

4.1 Multinomial Naïve Bayes

The Naïve Bayes is very simple as well as effective classification algorithm. The Naïve Bayes classification model is fast to build and it makes quick predictions. Naïve Bayes is a probabilistic classifier and it learns the probabilities of features based on the target class. It assumes that the occurrence of a particular attribute is independent of the occurrence of the other attributes. The multinomial Naive Bayes classifier is suitable for classification with discrete features like word counts for text classification. The multinomial distribution normally requires integer feature counts. This is especially useful when the whole dataset is too big to fit in memory at once. The hyper parameter alpha is tuned and log loss is measured.

4.2 Linear SVM

Data scientists often use Support Vector Machines (SVM), a powerful tool for classification due to their tendency not to overfit, but to perform well in a variety of problem domains. SVM performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. Mostly simple type SVM is applied on binary classification, dividing data points either in 1 or 0.

For multiclass classification, the same principle is utilized. The multiclass problem is broken down to multiple binary classification cases, which is also called one-vs-one. In scikit-learn one-vs-one is not default and needs to be selected explicitly. One-vs-rest is set as default. It basically divides the data points in class x and rest. Consecutively a certain class is distinguished from all other classes.

4.3 Logistic Regression

Logistic regression for multiclass classification follows the same ideas as the binary classification, used when the dependent variable is categorical. In the one vs rest method, when we work with a class, that class is denoted by 1 and the rest of the classes becomes 0. Logistic regression uses a sigmoid function to predict the output. The sigmoid function returns a value from 0 to 1. Generally, we take a threshold such as 0.5. If the

sigmoid function returns a value greater than or equal to 0.5, we take it as 1, and if the sigmoid function returns a value less than 0.5, we take it as 0.

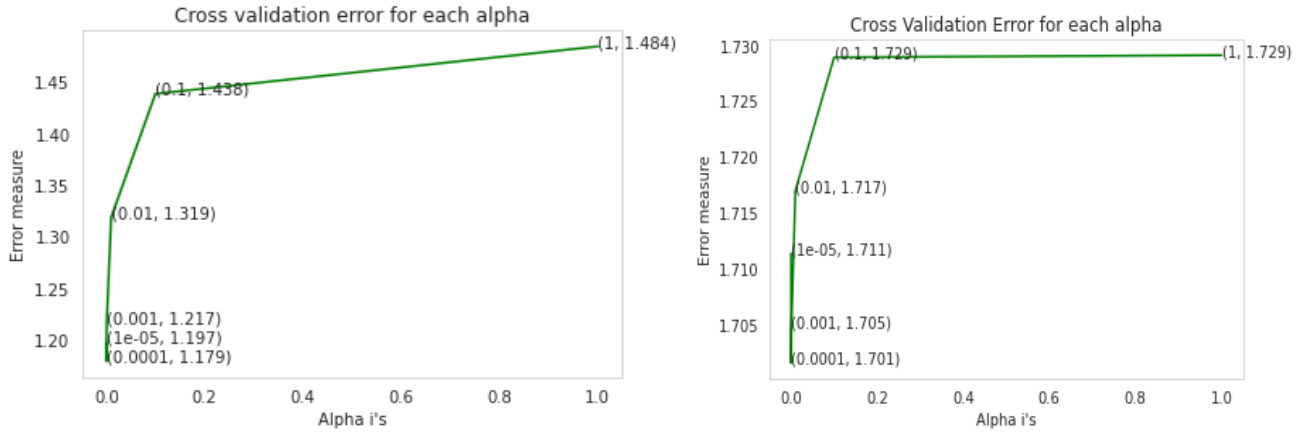
Multiclass Classification using Logistic regression is implemented using gradient_descent function. This function will take input variables, output variable, theta, alpha, and the number of epochs as the parameter. Here, alpha is the learning rate.

4.4 K Nearest Neighbours

K Nearest Neighbours (KNN) is a simple algorithm, which assumes that similar things are in close proximity of each other. So if a datapoint is near to another datapoint, it assumes that they both belong to similar classes. We successfully implemented a KNN algorithm for the msk dataset. We split the our input and output data into training and testing data, as to train the model on training data and testing model's accuracy on the testing model. We choose 80%–20% split for our training and testing data. Here, we see that the classifier chose 5 as the optimum number of nearest neighbours to classify the data best. Now that we have built the model, our final step is to visualise the results.

V. Results and Discussion

Multiclass classification is an effective machine learning application. In this article, the performance of machine learning algorithms Naïve Bayes, Linear Support Vector Mchine, Logistic regression and K Nearest Neighbor is evaluated.



(a) (b)

Figure 1.2. Measurement of log loss using SGD classifier
(a) Gene Expression (b) Variation Expression

The experimental results are also given for performance analysis of preprocessing by measuring log loss using SGD Classifier for the Gene, Variation and Text description are shown in table 1.1.

Table 1.1 Measurement of Log Loss of expression for preprocessing

Value of alpha	SGD Classifier		
	Gene	Variation	Text
0.00001	1.197163	1.711327	1.817415
0.0001	1.178994	1.701487	1.639308
0.001	1.216857	1.704789	1.383592
0.01	1.318580	1.716944	1.387875
0.1	1.438311	1.728920	1.252675
1.0	1.484302	1.729109	1.065219

After the knowledge of individual feature contribution to reduction in losses, all the three features are combined and train it with different models. Before subjecting the model to test data tuning of hyper parameters are done and the effects given by each value during tuning are displayed.

Evlauating performance of Multinomial Naïve Bayes classification model is based on conditional probability and uses Bayes theorem to predict the class of unknown datasets. This model is mostly used for large datasets as it is easy to build and is fast for both training and making predictions. MSK dataset is split into train and test-set for the following training and prediction. The train log loss value 0.9157884338872256, the CV log loss value 1.2169791659928892,the test log loss value 1.3294457054334512 is achieved for values of best alpha = 0.1.For The graphical representation is clearly given in the Figure 1.3.

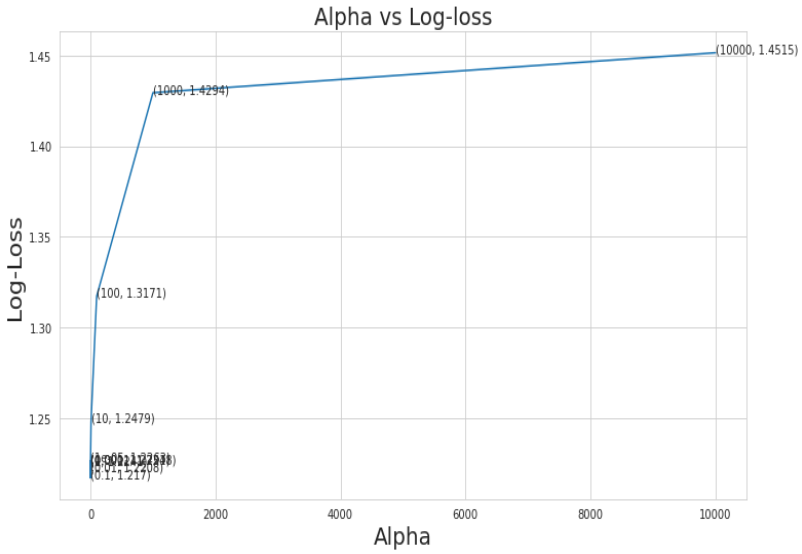


Figure 1.3 Measurement of Log Loss of expression using NB Classifier

Linear Support Vector Machine model is also subjected to tuning of hyper parameter alpha. For values of best alpha = 10, the log loss achieved for train data, CV data and test data are 0.7066571763065957, 1.180779798772505 and 1.2474154215717312 respectively. Percentage of Misclassification for CV points is 37.59% and for Test points, 39.85% is achieved.

Performance of Logistic regression is measured by tuning hyper parameter alpha. For values of best alpha = 10, the train log loss is 0.6259484564305449, the CV log loss is 1.055688362790151 and test log loss achieved is 1.132395511333499. Achieved percentage of Misclassification for CV points is 37.41% and for Test points, 37.74% is achieved.

Table 1.2 Measurement of Log Loss of expression for classification

Value of alpha	Measurement of Log Loss of Text expression		
	Multimonial NB	Logistic Regression	Linear SVM
0.00001	1.226251	1.830889	1.830889
0.0001	1.224789	1.597739	1.598945
0.001	1.225082	1.417703	1.418053
0.01	1.220826	1.421158	1.417248
0.1	1.216979	1.235759	1.416976
1.0	1.224119	1.072379	1.252674
10	1.247888	1.055549	1.168668
100	1.317129	1.184012	1.240891
1000	1.429444	1.418940	1.461083
10000	1.451518	1.504455	1.517734

For KNN, We calculate overall accuracy of the model by tuning hyper parameter Neighbor value. For values of best neighbors 181, the train log loss is 0.983631587810877, the CV log loss is 1.1685896121210864 and the test log loss is 1.1842975657550283. The graphical representation of cross validation error measurement is clearly shown in Figure 1.4.

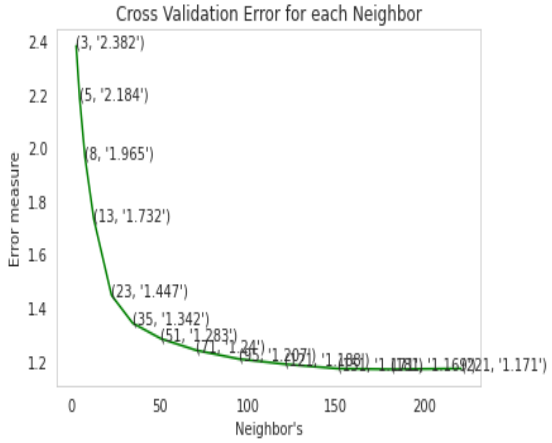


Figure 1.4 Measurement of Cross Validation Error using KNN

VI. CONCLUSION

Multiclass classification is an effective machine learning application used for classifying instances into one of three or more classes. Empirical studies are addressing that there are nine classes of cancer patients varied from high to low risk groups. Experimental results provides clinicians a general view of which prediction features are important in order to assign a patient to a specific clinical outcome. The objective of this study is to evaluate the performance of the multiclass classification algorithms for categorizing genetic mutations. SGD Classifier model is used for preprocessing and observing the contribution to the loss reduction. Naïve Bayes, Linear SVM, Logistic regression and KNN models are used as multiclass classifiers for genetic mutation. Through the result analysis, it is clearly stated that the performance of the classification is dependent on the machine learning algorithms techniques used. All these state of art models are mostly used for large datasets as it is easy to build and is fast for both training and making predictions. Future work can be directed towards evaluating different feature selection and prediction algorithms.

REFERENCES

[1] QingLiao, YeDing Zoe L.Jiang XuanWang ChunkaiZhang QianZhangMulti-task deep convolutional neural network for cancer diagnosis, Neurocomputing, Volume 348, 5 July 2019, Pages 66-73, https://doi.org/10.1016/j.neucom.2018.06.084.

[2] Khushboo Munir, Hassan Elahi, Afsheen Ayub, Fabrizio Frezza, and Antonello Rizzi, Cancer Diagnosis Using Deep Learning: A Bibliographic Review, Cancers (Basel). 2019 Sep; 11(9): 1235.

[3] Paul R., Hawkins S.H., Hall L.O., Goldgof D.B., Gillies R.J. Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT; Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC); Budapest, Hungary. 9–12 October 2016.

[4] Guo, Y.; Shang, X.; Li, Z. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing 2019, 324, 20–30.

[5] Golden, J.A. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen. JAMA 2017, 318, 2184–2186.

[6] Li, L.; Pan, X.; Zhang, L. Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. Multimed. Tools Appl. 2018, 810, 85–95.

[7] Zhu, Z.; Albadawy, E.; Saha, A.; Zhang, J.; Harowicz, M.R.; Mazurowski, M.A. Deep learning for identifying radiogenomic associations in breast cancer. Comput. Biol. Med. 2019, 109, 85–90.

[8] Bejnordi, B.E.; Veta, M.; Van Diest, P.J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.A.; Hermsen, M.; Manson, Q.F.; Balkenhol, M.; et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017, 318, 2199–2210.

[9] Bi, W.L,Hosny, A. Schabath, M.B.,Giger, M.L., et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J. Clin. 2019, 69, 127–157.

[10] Lamy, J.B.; Sekar, B.; Guezennec, G.; Bouaud, J.; Séroussi, B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artif. Intell. Med. 2019, 94, 42–53.

[11] Shen, L.; Margolies, L.R.; Rothstein, J.H.; Fluder, E.; McBride, R.; Sieh, W. Deep learning to improve breast cancer detection on screening mammography. Sci. Rep. 2019, 9, 1–12.

[12] Coccia, M. Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence. Technol. Soc. 2020, 60, 101198.

[13] Wang, H.; Yoon, S.W. Breast cancer prediction using data mining method. In Proceedings of the IIE Annual Conference Expo 2015, Nashville, TN, USA, 30 May–2 June 2015; pp. 818–828.

[14] Ahmad, L.G. Eshlaghy, A.Poorebrahimi, A, Ebrahimi, M, Razavi, A. Using three machine learning techniques for predicting breast cancer recurrence. J. Health Med. Inf. 2013, 4, 3.

[15] Mandal, S.K. Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree. Int. J. Eng. Comput. Sci. 2017, 6, 20388–20391.

[16] Borges, L.R. Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. Group 1989, 1, 369.

[17] Chaurasia, V.; Pal, S.; Tiwari, B. Prediction of benign and malignant breast cancer using data mining techniques. J. Algorithms Comput. Technol. 2018, 12, 119–126.

[18] Kumar, V.; Mishra, B.K.; Mazzara, M.; Verma, A. Prediction of Malignant & Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. arXiv 2019, arXiv:1902.03825.

[19] J.; Schmidt, M. Linear convergence and support vector identification of sequential minimal optimization. In Proceedings of the 10th NIPS Workshop on Optimization for Machine Learning, Long Beach, CA, USA, 8 December 2017; p. 5.

[20] Al-Sabbah, S.A.; Mohammad, S.F.; Eanad, M.M. Use of the Naive Bayes Function and the Models of Artificial Neural Networks to Classify Some Cancer Tumors. Indian J. Public Health Res. Dev. 2019, 10, 1563–1569.