A Study on the Performance of Classification Models for COVID-19 Datasets

Selvanayaki Kolandapalayam Shanmugam[a] and Shyamala Devi J[b]
[a]
Department of Computer Science, Concordia University Chicago, IL, USA
[b]Department of Computer Science and Applications, SRM Institute of Science and Technology, Chennai, India
[c]Assista
[d]Assist
[e]Assista
[f]Assistan

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: In today's world, classification algorithm is successfully applied to predict the zone based on the zip code of the dataset. Different classification algorithms can are applied to the dataset to devise method, which predicts the status of the zone for the given zip code available in the dataset. The accuracy of the methods differs from the classification algorithm used. It is very challenging to identify the best classification algorithm. Here, we attempt to make a comparative analysis and evaluations of the performance of different classification are done for the given data sets. As the datasets are getting increases, the algorithm performed better for its accuracy

Keywords: Classification, accuracy, precision, recall, dataset, F-Measure

1. Introduction

Machine Learning is the field of study and is relating to algorithms that learn from examples. In Machine learning, learning is classified as supervised learning and unsupervised learning.

Classification is a process which uses the Machine learning algorithms that learns how to assign class label to examples from any problem domain. There are different types of classification, in general, classification refers to a predictive modeling problem because class label is predicted for the given input data. To handle this approach, say, in modeling perspective, classification needs a training dataset of different set of input and outputs from which the system /model learns. Then, the model uses the training dataset and calculates how to best map the given input data to a specific class label. For constructing a classification model, for a problem many classification model types exist in today's world. Each type of classification model work differently for different dataset on similar domain and even different with different nature of same type of datasets. In general, researcher comes out with recommendations based on their experiments conducted and discovers the performance of different classification algorithms. The study explains or portrays that the one of the best metric for identifying the performance of model based on the predicted class labels is the classification accuracy. Though it is not a perfect justification, but a good start up. Moreover, the main types of classification identified are:

(i). Binary Classification – As name suggest, the classification has only two labels, 0 or 1 (i.e) Normal state or Abnormal state. To support this, algorithms used in this classification are Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), Naive Bayes (NB). Support Vector Machines (SVM) and Logistic Regression (LR) completely supports only binary classification.

(ii). Multi-Class Classification – This classification have two class labels. The classification is performed and belongs to one among range of class labels. The best example for multi-class classification is face classification. To support this, the algorithms used are KNN, Decision Trees, Naïve Bayes, Random Forest and Gradient Boosting.

In certain situations, classification are referred as in balanced, when the number of examples in each class is unequally distributed. (i.e.) majority of examples belongs to normal class and remaining examples belongs to abnormal class. Some examples are Credit card fraud detection, Medical diagnostic tests etc. The algorithms, SVM, Logistic Regression, Decision Trees are used for this method. The performance of this method is identified by its accuracy but it is reported that it is misleaded sometime, so, other performance metrics like precision, recall and F-Measure are used for classification where the examples in the dataset are unevenly distributed.

Coronavirus disease is a respiratory disease, which is caused by high acute respiratory syndrome. In December 2019, COVID-19 was detected in Wuhan, China and spreads worldwide and leads to coronavirus pandemic situation. More than, 4.2 million cases have been registered in more than 200 countries as on May 2020. The increase rate in India becomes higher and especially in Tamil Nadu, the rate of increase is higher during May-June 2020. The government has announced the status of risk in different places by dividing the state as Zones with districts and assign different colors highlighting the status of risk. As there are no vaccines to protect against COVID-19, the spread could be reduced by the isolation of the infected individuals.

The remaining part of the work is planned as follows: Literature review relating to this study is discussed in Section II. Section III describes the review of the classification models and techniques in the research. Section

IV comprises the results and discussion, and the Conclusion of the study is included in Section V.

2. Literature Study

Data Mining techniques is used to analyse the hidden information in larger dataset. Data Mining is defined as the process of discovering the messages hidden, patterns and knowledge with large datasets and results in giving predictions as outcomes. [5].

The different datamining techniques used in the life time applications are Pattern Recognition, Clustering, Association and Classification [2]. Among this, Classification is been noted as an important problem in the emerging field of data mining[3] , which finds the ways in interpreting the data sets. The most significant task is the classification of data in data mining, and achieves the goal of classification by predict the value of a designated discrete class variables [4].

Sa'di et. all analysed the data mining algorithms for PIDD and discussed that the accuracy achieved is 76.95% for Naïve Bayes classification algorithms [8]. For the same dataset, Seera and Lim proposed a hybrid classification method for Fuzzy Min-Max Neural Network, Regression Tree and Random forest and results the accuracy of 78.39% [7]. Roopa et. all built a model using diabetes data projected to a new space using Principal Component Analysis, then Linear Regression method is applied to obtain the accuracy of 82.1% [9].

Abdul Waheed et. all discussed the method to generate synthetic chest X-ray(CXR) images using ACGAN based model, CovidGAN. It is observed that the accuracy is increased to 95%. Mustafa Agaoglu, used different classification models for predicting the performance of the instructor in higher education. It is shown that classifier works best with respect to classification metrics. The Commonly used classification metrics are accuracy, precision and specificity.

Nora and Heba, used MERS-CoV dataset and examined the performance of classifier model for binary, multi-class and multi-label classification types and identified that the classifier, Decision Tree (DT) works best for binary classification with an estimated accuracy of 90% [6].

3. Methodology

In Machine Learning, one of the best supervised learning approach is classification, in which, computer programs learns from the data supplied to it and predicts new classifications or observations.

3.1 Classification Models

In Machine learning, classification model based algorithms are important domain. The major objective of classification is assigning class labels for the set of variables. This is done by a classifier model, where learning algorithm is applied on a training dataset. To understand effectively, it is needed to know the class label for each instance in the training dataset before training. Once the learning phase is done, the test set is used to evaluate the performance of classifier model. To handle this situation, different methods and algorithms for building the classifier mode are:

Logistic Regression:

Logistic Regression is a supervised Machine Learning approach. Logistic Regression is a technique, which has the base foundations of statistics. It uses the function, logistic function, called as sigmoid function developed by statisticians for describing the growth in ecology, rising quickly and maximizing the carry on capacity of the environment. It is a S-shaped curve, which takes any real-valued number and it will get mapped into a value in a range from 0 to 1. Logistic regression is a linear method and the predictions of this method are transformed using the logistic function.

Decision Trees:

It is a simple representation for performing the process of classification. Decision Tree (DT) is a supervised Machine Learning approach, where the data is split continuously in accordance with estimated parameter. It consists of Nodes, Edges, Leaf Nodes and is typed as Classification trees and Regression trees. According to the nature of "Fit" or "Unfit" results, the leaf nodes of the trees are classified under classification trees. Such building process of tree is known as binary recursive partitioning. In Decision Trees, the target variables takes continuous values, called as regression trees.

In Decision Tree Classifier algorithm, the process starts at the root of the tree, which splits the data based on the feature, which outcomes largest Information Gain(IG). The splitting procedure is repeated for each node till the leaves are pure. In reality, to prevent overfitting the limit is set for the depth of the tree. The advantage of using Decision Tree in classification are inexpensive to construct, Fast in classifying unknown records, easy in interpreting small sized trees, highly accurate for simple data sets. The limitations are decision boundary is restricted, showing biased natüre while spitting on features with larger number of levels, high impact on decision logic for a identified smaller changes in training data set, not merely accurate for complex data sets.

K-Nearest Neighbors:

The K-Nearest Neighbors (KNN) algorithm is a simple, flexible, easy-to-implement supervised machine learning algorithm helps to solve classification and regression problems. In this algorithm, it is assumed that similar things exist in close proximity. The idea of similarity is captured by the simple calculation of identfying the distance between the points on a graph. To handle this algorithm, steps included are,

1. Loading the data.
2. Intializing K to selected number of neighbors.

3. Calculate the distance between query value and current value for every data in the dataset.
4. Distance and index is added to the ordered collection.
5. Sorting the collection from smallest to largest.
6. Identify the labels, for noted 'K' entries and return the mode for the K labels.

The steps are iterated for the various value of "K", "K" value is decided for which the algorithm gives highest accuracy. The above depicts that KNN algorithm is simple, flexible and easy to implement. It is versatile, could be used for classification, regression and searching.

Naive Bayes

The classification algorithm based on Baye's theorem highlights the inclusion of assumption of independence among the predictors. It is classifier which assumes the particular feature presence not related to any other feature presence in a class.. It is easy in building and useful for the large data sets, identified to be a highly sophisticated classification methods. This algorithm works with steps, Converting the dataset into table, fiding the probabilities, Naïve Bayesian equation is used to calculate the posterior probability for each class. The class with highest posterior probability is considered as the prediction. Though NB classifier is easy and fast in predicting class, performs good in the catogorical input varibales, it is not recommended for real life applications where the independence exist with the set of predictors.

3.2 Performance Metrics

In Machine Learning, performance metrics are used for evaluating the algorithms used in classification problems. The classification metrics used are Log-Loss, Accuracy, AUC (Area under Curve). Apart from this, other evaluation metrics are Precision, recall. Relating to this, the easiest method for finding the correction and confusion matrix is the Confusion Matrix. Confusion matrix helps in representing the actual and predicted values in two dimensions. "Actual" classifications are columns and "Predicted" ones are Rows. In reality, perfromance metrics are based on Confusion Matrix. Confusion matrix has identifed cases which are True Positives, True Negatives, False Positives and False Negatives.

In performance metrics, Accuracy is a good measure when the target variable classes in the data are balanced. It is the number of correct predictions made by the model over all predictions identified.

$$Accuracy=TP+TN/(TP+FP+FN+TN)$$

Precision details about its performance in relating to false positives. It is all about Precise.

$$Precision=TP/(TP+FP)$$

Recall details the classifiers performance in relating to false negatives. The another performance metrics is Specificity, which is a exact opposite of Recall.

It is not necessary to highlight the Precision and Recall when we find the performance metrics for the classification problem. It is good to have a score which highlights these two, it is F1-Score. It is easily calculated by finding the arithmetic mean of Precision and Recall. But it doesn't work all times and so we apply harmonic mean to calculate F1 Score. So,

$$F1Score=2*Precision*Recall/(Precision + Recall).$$

4 Experimental Results & Discussion

The main focus of this study is to analysis and demonstrate the computational performance benefits as a metric for different Machine learning classification algorithms. In this section, we detail the implementation of different classification models for the COVID-19 related dataset.

4.1 Data Set Generation

The dataset is composed of 6 attributes discussing the details of the places with the impact of COVID-19 diseases. Then, class label is generated, identified as "Risk" and "Non-Risk" zone with class label values "1" or "0". We collected the data from data issued by government of Tamil Nadu. The decision to develop the dataset is taken by the fact that all of them are completely available to the public. The samples of dataset for binary classification are given in Figure 1 and Multiclass classification is given in Figure 2.

	Zone_label	Zone_name	Zone_subtype	zip	lat	long	color_score
0	1	Erode	Green	638001	11.33	77.72	1
1	1	Salem	Green	636001	11.66	78.16	1
2	1	Nammakkal	Green	637001	11.21	78.16	1
3	1	Karur	Green	639001	10.96	78.07	1
4	1	Thiruppur	Green	641604	11.10	77.34	1

Figure 1 : Dataset showing the samples of data – Binary Classification

	Label	District	Color-Zone	Pincode	latitude	longitude	Zone-Status	Risk
0	0	Erode	Green	638001	11.33	77.72	1	Low
1	0	Salem	Green	636001	11.66	78.16	1	Low
2	0	Nammakkal	Green	637001	11.21	78.16	1	Low
3	0	Karur	Green	639001	10.96	78.07	1	Low
4	0	Thiruppur	Green	641604	11.10	77.34	1	Low

Figure 2 : Dataset showing the samples of data – Multiclass Classification

4.2 Results and Discussion

The implementation of the study is done with Python3. Before applying the models to classify the data in the dataset, all preprocessing of data is done using standard libraries in Python. The performance of the classification models on this dataset is done by considering the metrics like Accuracy, Precision, Recall and F1-Score. The experimental results of Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees (DT) classifiers for binary classification datasets are shown in the table below. To compare the techniques fairly, same datasets are used.

Classification Models/Performance Metrics.	**Accuracy**	
	Training	**Test Set**
Logistic Regression(LR)	0.81	0.90
Naïve Bayes (NB)	1.00	1.00
K Nearest Neighbors (KNN)	0.93	0.90
Support Vector Machine (SVM)	0.78	0.90
Decision Trees (DT)	1.00	1.00

Table 1: Comparison of different classifiers using Binary Classification dataset.

Using the KNN- Classifier, the Risk status of the Zone is identified by given the user input of Pin code of the region. The System implemented accepts the pincode as a input, identifies the necessary longitude and latitude location of the region using Four Square API. The value of pincode, latitude, longitude are passed as parameters to predict the risk of the location in the given dataset.

```
Enter the Pincode for the search location: 638001
```

```
Enter the Pincode for the search location: 638001
The location 638001 with color status:
'Green'
```

The experimental results of different classifier models for the multiclass classification datasets are shown in the table below.

Classification Models/Performance Metrics.	**Accuracy**	
	Training	**Test Set**
Logistic Regression (LR)	0.67	0.50
Naïve Bayes (NB)	1.00	1.00
K-Nearest Neighbors (KNN)	0.81	0.80
Support Vector Machine (SVM)	0.41	0.40
Decision Trees (DT)	1.00	0.80

Table 2: Comparison of different classifiers using Multiclass Classification dataset.

The precision and Recall is discussed in section III – Methodology, and the experimental results are given below.

```
              precision    recall  f1-score   support

           0       0.90      1.00      0.95         9
           1       0.00      0.00      0.00         1

   micro avg       0.90      0.90      0.90        10
   macro avg       0.45      0.50      0.47        10
weighted avg       0.81      0.90      0.85        10
```

Figure: 3 : Confusion Matrix for Binary Classification Dataset.

	precision	recall	f1-score	support
1	0.25	1.00	0.40	1
2	0.00	0.00	0.00	2
3	0.00	0.00	0.00	1
4	0.75	1.00	0.86	3
5	0.50	0.50	0.50	2
6	0.00	0.00	0.00	1
micro avg	0.50	0.50	0.50	10
macro avg	0.25	0.42	0.29	10
weighted avg	0.35	0.50	0.40	10

Figure: 4 : Confusion Matrix for Multiclass Classification Dataset.

From the above analysis, it is identified and evident that the classifiers work better for various classifiers and increases the performance if the number of attributes used in the dataset are getting increased.

5. Conclusion

In this machine learning base learning model for identifying the location and its risk status could be used for any type of data driven smart applications in intelligently in finding the results depends on the values given. Here, the dataset is build for binary classification, for which the performance metrics is analysed for different classification models. The same is been represented in graph format also. The data can be interpreted as multiclass representation, so, the same identified dataset is explored in different perspectives and built the multiclass classification dataset. The performance metrics for this multiclass datasets are experimented and analysed results is given above with graph representation. It is believed that the effectiveness analysis on various machine learning classification will help researchers in this field to appropriately select the model and be used as reference. In future, identification and detection of location details with different characteristics is done using the best classifier model discussed in this work.

References

1. Gopala Krishna Murthy Nookala., Bharath Kumar Pottumuthu., Nagaraju Orsu., Suresh B.Mudunuri.: Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification. In: International Journal of Advanced Research in Artificial Intelligence, vol. 2, no.5, pp. 49-55.(2013)
2. Han, J., and Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, (2006)
3. Chen, M.S., Han, J., Yu, P.S.: Data mining: an overview from a database perspective. In: IEEE Transactions on Knowledge and Data Engineering, vol. 8, no.6, pp. 866 – 883, (2002)
4. Grossman, D., Domingos, P.: Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood. In: Proceedings of the 21st International Conference on Machine Learning, Banff, Canada (2004)
5. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Waltham, MA, USA: Morgan .Kaufmann (2012)
6. AlMansour, Nora., Kurdia, Heba.: Identifying Accurate Classifier Models for a Text based MERS-CoV Dataset. In: Intelligent Systems Conference, pp.430-435 (2017)
7. Raja, K.S., Kiruthika, U. An Energy Efficient Method for Secure and Reliable Data Transmission in Wireless Body Area Networks Using RelAODV. Wireless Pers Commun 83, 2975–2997 (2015). https://doi.org/10.1007/s11277-015-2577-x
8. Seera, M., Lim, C. P.: A hybrid intelligent system for medical data classification. In: Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249 (2014)
9. Sa'di, S.A., R.Hashemi, Maleki., Panbechi, Z., Chalabi, K.: Comparison of data mining algorithms in the diagnosis of type II diabetes. In: International Journal of Computer Science Applications , vol. 5, no. 5, pp. 1–12 (2015)
10. Roopa, H., Asha, T.: A linear Model based on Principal component Analysis for Disease Predication, Vol: 7, pp. 105314-105318 (2019)
11. Abaidullah, A. M., Ahmed, N., Ali, E.: Identifying hidden patterns in students' feedback through cluster analysis. In: Int. J. Comput. Theory Eng., vol. 7, no. 1, pp. 16–20 (2015)
12. Delavari, N., Phon-Amnuaisuk, S., Beikzadeh, M. R.: Data mining application in higher learning institutions. In: Inform. Edu.-Int. J., vol. 7, no. 1, pp. 31–54 (2007)
13. Kentli, F. D., Sahin, Y.: An SVM approach to predict student perfor- mance in manufacturing processes course. In: Energy, Edu., Sci. Technol. B, vol. 3, no. 4, pp. 535–544 (2011)