Implementing Different Classification Techniques to predict the Wine types

Rabin Chhetri - 001185145

Abstract

Different Classification Techniques like Logistic Regression (LR), Support Vector Machine (SVM) and K Nearest Neighbour(KNN)are successfully implemented over the Wine dataset to classify the types of wine based on 178 samples with 13 features like alcohol, malic acid,phenols etc.The number of feature variables are reduced to 2 and model is prepared to classify the type of wine. After Successful implementation of three different methods, Support vector Machine is found to be more accurate in classifying the wine type.

1. Introduction

In todays World, Wine has become an integral part of people's lifestyle(Garima Agrawal 2018). From teenagers to old aged people most of them are found enjoying their quality time with friends, families and colleagues with a glass of wine.Wine is composed of different materials so its taste depends upon the qunatity and quality of its composition so this document is prepared to classify the type of wine based on its composition materials.To classify a wine into three different types, I have used classification techniques like Logistic Regression (LR), Support Vector Machine (SVM) and K Nearest Neighbour (KNN) algorithms.

Classification is one of the major activity that is pravailing in the world. different Classification methods are used to classify the various aspects of any product. Wine being liquid material it is widely consumed by the people in their daily life. The quality of the wine directly affects the health of the people.With an aim to ease the consumer in identifying the type of wine based on its composition, different researchers have implemented various algorithms to identify the wine type like ART1 Network(N. C. Kavuri 2011) and many more.

This document also highlights on some classification approaches. I have taken the Wine dataset using the scikit learn library. The dataset has 178 sample data with 13 feature variables like alcohol, malic acid, phenols,ash, magnesium etc and a target variable. Initially, the number of feature variable is reduced to 2 from 13 and only the first two features alcohol and malic acid are used to predict the target variable. Out of 178 samples, i have used 100 samples to train the model and remaining 78 samples to test whether the model created is giving the desired prediction or not. I have used above mentioned three classification methods to train and test the model using the dataset. various evaluation metrices like accuracy, confusion matrix, precision, recall, f1-score and support were analysed after implementing the classification methods.

2. Methods

This part of the report contains the detail description of the classification algorithms that has been implemented on solving this wine classification problem.The Used algorithms are Logistic Regression (LR), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN).

2.1. Logistic Regression (LR)

Logistic Regression(Stephan dreiseit 2002) is the widely used statistical model for classification purposes . For a given a set of Sample data, Logistic Regression algorithm helps us to classify these observed values into two or more discrete classes. So, the nature of the target variable is discrete.

In the following, if x,z denotes the feature and the target variables respectively then the linear equation would be given mathematically by

$$z = \beta_0 + \beta_1 x_1 \tag{1}$$

where the coefficients β_0 and β_1 are the parameters of the model

So if there are multiple feature variables then the above equation 1 can be extended as

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n \tag{2}$$

here z is the target variable and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_n$ are the parameters of the logistic regression model. This target variable z is now converted into probability value which lies between 0 and 1. Sigmoid function is used to convert the predicted value to the probability value. This sigmoid function is given by the below equation.

$$\phi(z) = \frac{1}{1 + e^{-}z} \tag{3}$$

2.2. Support Vector Machine (SVM)

Support vector Machine (SVM) are a set of supervised machine learning algorithm used for classification, regression and outliers detection.This method is effective in a high dimensional spaces(Paul Pavlidis 2004). It transforms the data in high dimensional space to change a linearly inseparable dataset to a linearly separable dataset because of which it is considered memory efficient.This algorithm is more accurate in small dataset and is prone to over fitting issue.
SVM works by finding a boundary which separates values from each other. In 2D space, the boundary is called line, In 3D it is called plane and in higher dimesions it is called hyperplane.

$$h(x_i) = 1 if w.x + b >= 0 \quad (4)$$

$$h(x_i) = -1 if w.x + b < 0 \quad (5)$$

here, wx + b =0 is defined as the hyperplane.
equation (4) says it is on or above the hyperplane so the class is +1 whereas equation (5) says it is below the hyperplane so the class is -1.
SVM uses the below equation to transform the data into the higher dimensional space.

$$\phi(x) = [x, x^2] \quad (6)$$

where x, x^2 represents the data in the lower dimensions and $\phi(x)$ represents data in higher dimensional space.

2.3. K Nearest Neighbour (KNN)

K Nearest Neighbour(Wang and Zhao 2012) is also one of the supervised learning algorithm. In this algorithm, for a set of variables x,y where x is the feature variable and y is the target variable. a function h:x $\rightarrow y$ is defined to find the realation between x and y so that for any unknown values of x the algorithm can easily predict the valus for y.
In this KNN classification algorithm, K nearest Neighbours of unknown data point is found and then assign the class to unknown data point by having the class which has the highest number of data points out of all classes of K neighbours.The distance between two data points p and q is calculated using eucledian distance(Li et al. 2015) formula as

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (7)$$

A low value of K causes noise in the data set which is also known as overfitting.

3. Experiments

This section of the report highlights on the feature varibales of wine dataset used, different evaluation criteria that are considered like accuracy, confusion matrix, f1 score and their comparative results between different methods

3.1. Experimental settings

Classification Methods described in section 2 are implemented on the wine dataset. this dataset is taken using inbuilt sklearn library. The experiment is carried out using google colab. Python libraries like numby, sklearn, pandas, matplotlib and seaborn are used to import and manipulate the dataset to properly fit in the above models. The Wine Dataset contains 13 feature variables. some of them are alcohol, ash, phenols etc. Out of 13 feature variables I have reduced the feature variables to just 2 so that it will be easier to fit into the model. Alcohol and Malic acid are the only two features which are used to train the model. this dataset contains 178 sample records out of which 100 samples are used as a training data set and rest 78 samples are used as a testing data set.After this setup, different models are created using Logistic Regression algorithm, Support vector machine algorithm and K nearest neighbours algorithm with the value of k as 4.

3.2. Evaluation criteria

After implementing the above algorithms on the wine dataset, different evaluation metrices are observed on the predicted data so know on how accurate the model can classify the new data. The below sub section decribes the different evaluation metrices used in the experiment to test the correctness of the trained model using classification algorithms.

3.2.1. Accuracy

Accuracy gives the fraction of number of total predictions which are correctly predicted by the model. It can have value from minimum 0 to maximum 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Here,
TP= True Positive
TN= True Negative
FP= False Positive
FN= False Negative

3.2.2. Precision

Precision gives the total number of fraction of correctly predicted positive values out of all predicted positive values.

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

3.2.3. Recall

Recall gives the total number of fraction of the correctly predicted positive values out of all positive values. It is also

known as Sensitivity of the model.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

3.2.4. F1-SCORE

F1 Score gives the harmonic mean of precision and recall from the model.It is considered as a better option to evaluate classification performance on imbalanced data.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

3.3. Results

This section contains some values that are obtained after successfully implementing the above methods in the wine dataset.

	LR	SVM	KNN
Accuracy	0.7949	0.8333	0.7949

Table 1. Accuracy of LR, SVM and KNN Methods

	Precision	Recall	F1 Score	Support
-1	0.64	0.78	0.70	18
0	0.93	0.79	0.85	33
1	0.79	0.81	0.80	27
Accuracy			0.79	78
Macro Avg	0.78	0.79	0.78	78
Weigh Avg	0.81	0.79	0.80	78

Table 2. Classification Report for LR

	Precision	Recall	F1 Score	Support
-1	0.68	0.72	0.70	18
0	0.91	0.88	0.89	33
1	0.85	0.85	0.85	27
Accuracy			0.83	78
Macro Avg	0.81	0.82	0.82	78
Weigh Avg	0.84	0.83	0.83	78

Table 3. Classification Report for SVM

These figures available in above seven tables are the experimental results obtained from all three classification methods namely Logistic Regression, support Vector Machines and K nearest neighbour. Based on the data from Table 1 it is clearly understood that the Support Vector Machine is found to be more accurate as compared to other two methods for classifying the wine using wine dataset.

	Precision	Recall	F1 Score	Support
-1	0.64	0.78	0.70	18
0	0.93	0.79	0.85	33
1	0.79	0.81	0.80	27
Accuracy			0.79	78
Macro Avg	0.78	0.79	0.78	78
Weigh Avg	0.81	0.79	0.80	78

Table 4. Classification Report for KNN

	Class A	Class B	Class C
Class A	0.17948718	0.02564103	0.02564103
Class B	0.03846154	0.33333333	0.05128205
Class C	0.06410256	0.00000000	0.28205128

Table 5. Confusion Matrix for LR

	Class A	Class B	Class C
Class A	0.16666667	0.03846154	0.02564103
Class B	0.02564103	0.37179487	0.02564103
Class C	0.05128205	0.00000000	0.29487179

Table 6. Confusion Matrix for SVM

	Class A	Class B	Class C
Class A	0.17948718	0.02564103	0.02564103
Class B	0.03846154	0.33333333	0.05128205
Class C	0.06410256	0.00000000	0.28205128

Table 7. Confusion Matrix for KNN

3.4. Discussion

Based on the experimental result in section 3.3, Support Vector Machine seems to be more efficient for this wine dataset problem. This Dataset has 13 feature variables that means it is high dimensional dataset so the SVM has a feature of Kernelling the Data from lower dimension to Higher dimensional space because of which it makes it easier to find the separator and classification becomes easier and accuracy is high.Also, KNN is implemented with only 4 nearest neighbours so the predicted model is with noise. it could give further more accurate predictions if the value of K is properly managed.

4. Conclusion

Hence, after successfull implementation of Logistic Regression, Support Vector Machine and K nearest neighbour algorithm to classify the type of wine based on its two compostion materials namely alchol and malic acid, It is concluded that SVM can classify more accurately based on the information available from the wine dataset. Also, other modern and efficient Machine learning algorithms can be implemented over this dataset to obtain more accurate re-

sults. Further, It is also concluded that reducing the number of feature varibales in the dataset helps in increasing the performance of the model.

References

Garima Agrawal, Dae-Ki Kang (2018). "Wine Quality Classification with Multilayer Perceptron". In: *International Journal of Internet, Broadcasting and Communication* 10.2, pp. 25–30.

Li, Zhang et al. (2015). "Weigted-KNN and its application on UCI". In: *2015 IEEE International Conference on Information and Automation*, pp. 1748–1750. DOI: 10.1109/ICInfA.2015.7279570.

N. C. Kavuri, Madhusree Kundu (2011). "ART1 Network: Application in Wine Classification". In: *International Journal of Chemical Engineering and Applications* 2.3, pp. 189–195.

Paul Pavlidis Ilan Wapinski, William Stafford Noble (Mar. 2004). "Support vector machine classification on the web". In: *Bioinformatics* 20.4, pp. 586–587.

Stephan dreiseit, Lucila ohno-Machado (2002). "Logistic regression and artificial neural network classificationmodels: a methodology review". In: *Journal of Biomedical Informatics* 35, pp. 252–259.

Wang, Lijun and Xiqing Zhao (2012). "Improved KNN classification algorithms research in text categorization". In: *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, pp. 1848–1852. DOI: 10.1109/CECNet.2012.6201850.