Transactions of the Nigerian Association of Mathematical Physics
Volume 15, (April –June, 2021 Issue), pp43 – 46

A COMPARISON OF CLASSIFICATION ALGORITHMS ON STUDENTS' PERFORMANCE IN COMPUTER SCIENCE PROGRAMME USING WEKA

[1] S. A. Ibrahim, [2] K.S. Isiaka and [3] I.O. Mustapha

[1]Department of Physical Sciences, Al-Hikmah University, Ilorin, Nigeria
[2]Department of Mathematics, School of Science, Kwara State College of Education, Ilorin
[3] Department of Computer Sciences, Al-Hikmah University, Ilorin, Nigeria

Abstract

In this study, an evaluation of two classification algorithms such as Logistic Regression Model Classifier (LRMC) and Supporting Vector Machine Classifier Linear Model (SVM) on students' performance in computer science programme was performed through Waikato Environment for Knowledge Analysis (WEKA). Data used for the study was obtained through secondary source from the Department of Physical Sciences, Al-Hikmah University Ilorin-Nigeria, on students' academic performance in computer science programme.

The results showed that logistic regression model classifier was highly efficient compared to SVM. It showed a better performance than SVM with 94.77% to 85.36% Accuracy. Hence, it can be stated that logistic regression model classifier can be use to build apredictive model for students' performance in computer science programme, because it was correctly labeled more students' with the minimum university admission requirement for M.Sc. through their B.Sc. obtained in Computer Science programme with little error rate compare to SVM.

Keywords: Classification Algorithms, Cross Validation, Students' performance, Metric Performance,WEKA

1.0 Introduction

Numerous educational research problems call for building a predictive model using Supervise–Classification Algorithms for a class label (or binary outcome). For instance, class label could be, whether an under graduate student is likely to spiel before graduating or not, whether a graduate students will satisfy minimum university admission requirements for Post Graduate Degree (e.g. M.Sc.) through their first degree (e.g. B.Sc.) or not and so on. Supervise–Classification Algorithms in Machine Learning such as Logistic regression model classifier (LRMC) and Support Vector Machine Classifier Linear Model (SVM) enable Data Scientist to build predictive model using these algorithms and data which have already been collected. For detail about Machine learning Algorithms see [1-5], among others. Implementation of predictive model is usually done on software.

Thus, the implementation of predictive models build for this study was carried out on Waikato Environment for Knowledge Analysis (WEKA). It was developed at the University of Waikato, New Zealand, It is a free software licensed under the GNU General Public License, and the companion software to the book "Data Mining: Practical Machine Learning Tools and Techniques". It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. It is used in many different application areas, such as Biology, Economics, Medicine, Education, among others area of research [1, 2, 6].

This study is aimed at evaluating metric performances of two Classification Algorithms, LRMC and SVM that model through 10-folds cross validation in WEKA for class label $Y_i\,(1,0)$, (i.e. yes or no), $if\ Y_i \in (2^2, 2^1, 1) = 1$ or $if\ Y_i \in (pass\ , 3^{rd}) = 0$. Where$Y_i \in (2^2, 2^1, 1)$ represent the students with the minimum University admission requirement for second degree (M.Sc.) through their first degree (B.Sc.) obtained in Computer Science programme in the Department of Physical Sciences, and $Y_i \in (pass\ , 3^{rd})$ represent the students without the minimum University admission requirement for second degree (M.Sc.) through their first degree (B.Sc.) obtained in Computer Science programme in the Department of Physical Sciences, together with mixture of continuous and class label features (predictor variables). The rest of this study is organized as follow: Section 2, described materials and methods used. Section 3, deals with results and discussion and Section 4, deals with conclusions.

Correspondence Author: Iabrahim S.A., Email:adesinas2010@alhikmah.edu.ng, Tel: +2348052262175

2.0 Materials and methods

2.1 Materials

2.1.1 Data Source

Data used for this study was obtained throughsecondary source from the Department of Physical Sciences Al-Hikmah University Ilorin-Nigeria, from 2009 to 2015 on students' academic performance in computer science program**,** with 478 instances (Graduating Students spreadsheets file). Thus, the following are the attributes included in the data collected**:** Age, state of origin, gender, cumulative grade point average (cgpa), total credits passed (tcp), mode of entry and class of degree for the various graduate students. Thus, 10-fold cross validation was used to model and evaluate the performance of classification models used.

2.2 Methods

Brief descriptions of the mathematical development of classification models (methods) used are provided in what follow:

2.2.1 Logistic regression Model for classification

Consider a collection of k-categorical or continuous features (or predictor variables) be denoted by vector $X^1 = (x_1, x_2, \dots x_k)$. let the conditional probability that the label class is present be denoted by $P(Y = 1|x_1, x_2, \dots x_k)$, then the logit or log odds of having $Y = 1$ is modeled as a linear function of features (or predictor variables) as:

$$ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad [7, 8] \quad (1)$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \quad (2)$$

$$p = \frac{1}{1+e^{-z}} \quad (3)$$

where$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (4)

andβ_0 is the constant or intercept and $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients $x_1, x_2, \dots x_k$ of respectively.

Thus, the decision boundary for two-class logistic regression lies where the prediction probability is 0.5. i.e.

$$p(Y = 1|x_1, x_2, \dots x_k) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} = 0.5 \quad (5)$$

This occurs when

$$-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k = 0 \quad (6)$$

Because this is a linear equality in the attribute values, the boundary is a plane, or hyper-plane, in an instance space. It is easy to visualize sets of points that cannot be separated by a single hyper-plane, and these cannot be discriminated correctly by logistic regression[1, 2].

2.2.2 Support Vector Machine (SVM) for classification

Given $(\boldsymbol{x_i}, \boldsymbol{y_i})$, $i = 1,2,3, \dots, n$ where x_i- input variable; y_i- output variable (label class), where the $\boldsymbol{y_i}$are either +1 or -1, each indicating the class to which the point $\boldsymbol{x_i}$belong. Each $\boldsymbol{x_i}$is a P- dimensional real vector.

We want to find the "maximum-margin hyper-plane" that divides the group of points $x_i$for which $y_i = +1$from the group of points for which$y_i = -1$,which is defined so that the distance between the hyper-plane and the nearest point x_i from either group is maximized. Any hyper-plane can be written as the set of points x satisfying

$$w^T x - b = 0 \quad (7)$$

where$w$ is the normal vector to the hyper-plane. This is much like Hesse normal form, except that w is not necessarily a unit vector. The parameter $\frac{b}{||w||}$determines the offset of the hyper-plane from the origin along the normal vector w[9]

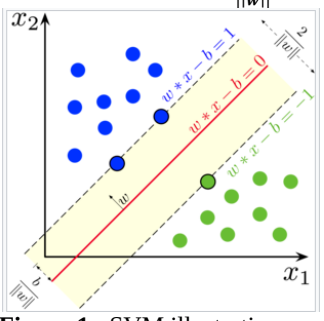

Figure 1: SVM illustration

Source : [9]

2.2.2.1 Hard margin

If the training data is linearly separable, we can select two parallel hyper-planes that separate the two classes of data, so that the distance between them is as large as possible. The region bounded by these two hyper-planes is called the "margin", and the maximum-margin hyper-plane is the hyper-plane that lies halfway between them. With a normalized or standardized dataset, these hyper-planes can be described by the equations

$w^T x - b = +1$ and (8)

$w^T x - b = -1$ (9)

Geometrically, the distance between these two hyper-planes is $\frac{2}{||w||}$. Thus, to maximize the distance between the planes we want to minimize $||w||$. The distance is computed using the distance from a point to a plane equation. We also have to prevent data points from falling into the margin, we add the following constraint: for each $\boldsymbol{i}$ either.

$$\boldsymbol{y_i} \begin{cases} +1 \quad if \quad w^T x - b \geq +1 \\ \\ -1 \quad if \quad w^T x - b \leq +1 \end{cases} \tag{10}$$

These constraints state that each data point must lie on the correct side of the margin. This can be rewritten as

$\boldsymbol{y_i}(w^T x - b) \geq +1$ for all $1 \leq i \leq n$ (11)

This can be put together to get the optimization problem as follow:

Minimize $||w||$ subject to $\boldsymbol{y_i}(w^T x - b) \geq +1$ for $i = 1,2,3,\dots,n$ (12)

The $\boldsymbol{w}$and$\boldsymbol{b}$that solve this problem determine our classifier, $\boldsymbol{x}$ →sign ($\boldsymbol{w^T x - b}$) where sign (.) is the sign function. An important consequence of this geometric description is that the max-margin hyperplane is completely determined by $\overrightarrow{\boldsymbol{x_i}}$ those that lie nearest to it. These $\boldsymbol{x_i}$are called support vectors [9].

2.3 Classifier's Performance Matrices

The results were presented using the following performance matrices of classifiers: Sensitivity, Specificity, Accuracy, Precision, and Time taken to build the model. The terms are defined as follow:

Accuracy = $\frac{(\text{TP+TN})}{(\text{TP+TN+FP+FN})}$ (13)

Precision = $\frac{TP}{(TP+FP)}$ (14)

Recall = $\frac{TP}{(TP+FN)}$ (15)

Specificity = $\frac{TN}{(TN+FP)}$ (16)

where, True positives (TP): when outcome is correctly classified to be positive (yes), when it is actually positive (yes). i.e. outcome were predicted to be a success and were actually observed to be success. False positives (FP): when outcome is incorrectly classified as positive (yes), when it is actually negative (no). i.e. outcome were predicted to be success but were actually observed to be failure. True Negatives (TN): when outcome is correctly classified to be negative (no), when it is actually negative (no), i.e. outcome were predicted to be a failure and were actually observed to be a failure. False Negatives (FN): when outcome is incorrectly classified as negatives (no), when it is actually positive, i.e. outcome were predicted to be a failure but were actually observed to be a success [2, 4, 5, 7].

3 Results and Discussion

The data of this study was analyzed using Waikato Environment for Knowledge Analysis. 10-folds cross validation was used in WEKA to model and evaluate the LRCM and SVM on students' performance in computer science programme. The process was done through partitioning the data into ten different folds. By this, $(K-1)$ *or* (9/10) of the data was used to train the model each time and the remaining K *or* (1/9) was used to test the model each time as well. Thus, the confusion matrices in Table 1-2 were obtained as well as model's performance in Table 3 for each classification algorithms as well as metric performance.

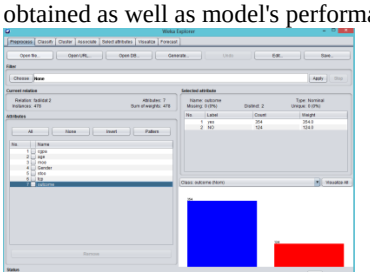

Figure 2: WEKA with explorer window open with study data contain 478 instances, while outcome variable contain yes (354) and no (124)

Table 1: Confusion matrix for logistic regression classification model

	Predicted model			
outcome		yes	no	Total
	yes	343	11	
	no	14	110	
	Total			

Table 2: Confusion matrix for SVM Linear classification model

	Predicted model			
outcome		yes	no	Total
	yes	324	30	
	no	40	84	
	Total			

Table 3: Metrics performance of logistic regression model and SVM

classifiers	Accuracy	Error rate	Precision	Recall	f-measure	Time taken
Logistic regression	94.77%	5.23%	96.10%	96.90%	96.50%	0.23sec
Linear SVM	85.36%	14.64%	89.00%	91.50%	90.30%	0.38sec

From accuracy point of view, it was fund that logistic regression model classifier has 94.77% which was better than that of SVM with 85.36%. Logistic regression classifier was correctly labeled more students' with the minimum university admission requirement for M.Sc. through their B.Sc. obtained in Computer Science programme to the whole pool of students under study than SVM, as confirmed from Table 3. Also, error rate of logistic regression model classifier was low compared to SVM, this due to high Accuracy that logistic regression model classifier possessed.

From precision, logistic regression model classifier has 96.10% which was better than that of Linear SVM with 89.00% in term of precision, this indicated that how many of those students that label with the minimum university admission requirement for M.Sc. through their B.Sc. obtained in Computer Science programme were actually with it, from the two classification Algorithms, as confirmed from Table 3.

From recall, logistic regression model classifier has 96.90% which was better than that of SVM with 91.50% in term of recall, this indicated that how many of those students that label with the minimum university admission requirement for M.Sc. through their B.Sc. obtained in Computer Science programme were correctly predicted, from the two classifiers, as confirmed from Table 3. Finally, the time taken to build the model was fewer in case of logistic regression model classifier as compare with SVM classifier.

4.0 Conclusion

In this study, two classification algorithms (LRMC and SVM) were successfully implemented for students' performance in computer science programme on WEKA. For the purpose of finding, LRMC was highly efficient compared to SVM. LRMC showed a better performance than SVM with 94.77% to 85.36% accuracy. Hence, it can be stated that logistic regression model classifier can be use to build a predictive model for students' performance in computer science programme because this classifier was correctly labeled more students' with the minimum university admission requirement for M.Sc. through their B.Sc. obtained in Computer Science programme with little error rate compare to SVM. In future study, LRMC can be compared with other classification algorithms.

References

[1] Witten, I.H. and E. Frank, *DATA MINING Pratical Machine Learning Tools and Techniques.* Elsevier Inc. , 2005(Second Edition).

[2] Witten, I.H., et al., *Data Mining Practical Machine Learning Tools and Techniques*Elsevier Inc., 2017(Fourth Edition).

[3] Brownlee, J., *Master Machine Learning Algorithms: discover how they work and implement them from scratch*. 2016: Machine Learning Mastery.

[4] Lantz, B., *Machine learning with R: expert techniques for predictive modeling*. 2019: Packt publishing ltd.

[5] Hackeling, G., *Mastering Machine Learning with scikit-learn*. 2017: Packt Publishing Ltd.

[6] Weka, *Wikipedia, the free encyclopedia.* 2021 July, 22(https://en.wikipedia.org/wiki/Weka_(machine_learning)).

[7] Ibrahim, S.A. and N.A. Fadil, *A Binary Logistic Regression Analysis of Students' Performance in Computer Science Programme.* Transaction of the Nigerian Association of Mathematical Physics, 2020. **12**(1).

[8] Peng;, C.J., K.L. Lee;, and G.M. Ingersoll, *An Introduction to Logistic Regression Analysis and Reporting.* The Journal of Educational Research, 2002. **96**(1).

[9] SVM, *Wikipedia, the free encyclopedia.* 2021 July,19(https://en.wikipedia.org/wiki/Support-vector_machine).