ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse66912020.pdf
https://doi.org/10.30534/ijatcse/2020/66912020

One-Class Conditional Anomaly Detection Algorithm (OCCADA) for Multiple Linear and Logistic Regression

Ivy Kim D. Machica[1], Bobby D. Gerardo[2], Ruji P. Medina[3]
[1]Technological Institute of the Philippines, Philippines, ikmachica@usep.edu.ph
[2]West Visayas State University, Philippines, bgerardo@wvsu.edu.ph
[3]Technological Institute of the Philippines, Philippines, ruji.medina@tip.edu.ph

ABSTRACT

Traditional one-class anomaly detection is error-prone, leading to numerous false-positive and false-negative anomalies within a context or condition. These anomalies or outliers tend to decrease the accuracy of the Multiple Linear Regression (MLR) and Logistic Regression (LR) models. Also, many real application scenarios have datasets consisting of normal data points or events without anomalies. This study introduces an algorithm using semi-supervised learning on a one-class dataset for classifying conditional anomalous instances. Secondary ground-truth datasets were collected from the Philippine Atmospheric, Geophysical, and Astronomical Services Region XI in Davao City, Philippines Natural Environment Research Council, and Kaggle repositories. The study used six (6) models for evaluation of the One-Class Conditional Anomaly Detection Algorithm (OCCADA) and its application to MLR & LR. The basic assumptions of the models were considered in making valid inferences. The results showed that OCCADA's classification accuracy on continuous and dichotomous response variables were 90% and 91%, respectively. Also, the application of OCCADA to MLR generated a low MAPE value of 0.09 compared to the classic MLR. Similarly, the application of OCCADA to LR increased in percentage of 18.30% compared to the classic LR. The results show that the use of the new one-class conditional anomaly detection algorithm using semi-supervised learning was effective in producing a highly accurate model for classifying conditional anomalous instances and improved the prediction accuracy of MLR and LR models.

Key words : Classification, Conditional Anomaly, One-Class, Multiple Linear, and Logistic Regression.

1. INTRODUCTION

Data mining is used to discover patterns using descriptive and predictive methods. The descriptive methods used in data mining are an association, clustering, summarization, and sequence analysis while predictive methods are classification, prediction, time series, and regression (Dunham, 2002; Patel & Mehta, 2011; Rezig, Achour, & Rezg, 2018; Sagar, Prinima, & Indu, 2017)

Predictive methods such as classification algorithms generate models from training samples $x_1,\ldots,x_n$ and using the learned model to classify and predict the new dataset. The algorithms for classification include Iterative Dichotomiser 3 (ID3), C4.5, Bayesian Network, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) among others. Moreover, regression predicts the real-valued data item by some known type of function (Dunham, 2002). The regression techniques are linear & multiple regression, logistic, lasso, and ridge, which are used to make predictions (Geetha, 2018).

Anomaly is considered as a special kind of outlier that is of interest with an analyst (Aggarwal & Heights, 2016). The three (3) types of anomalies are point, contextual or conditional, and collective (Goldstein & Uchida, 2016). The supervised, semi-supervised, and unsupervised learning methods can be used to detect the anomalies. The algorithms used to identify point & collective anomalies using supervised learning are KNN, Local Outlier Factor (LOF), and Distributed Time-Delay Neural Network (DTDNN). The One-Class Support Vector Machine (OCSVM), Feature Boundaries Detector for One-Class Classification (FBDOCC), Least Squares OCSVM (LS OCSVM), Kernel Principal Component Analysis (PCA), Gaussian Process One-Class (GP OCC), Condensed Nearest Neighbor Data Description (CNDD), and One-Class Random Forest (OCRF) are used for semi-supervised learning. Also, Unsupervised Principal Component Classifier (UNPCC) and DTDNN can be used in multi-class dataset for unsupervised learning. The Multivariate Conditional Outlier Detection (MCODE) was developed for multi-class and supervised learning while One-Class Conditional Random Fields (OCCRF) for unsupervised learning. However, there are no existing algorithms for semi-supervised learning for a one-class dataset for conditional anomaly detection to date.

Moreover, traditional one-class conditional anomaly detection is error-prone, leading to numerous false-positive and false-negative errors. One of the cases of false-positive reading is in diagnosing heart conditions when physicians

misdiagnose a healthy patient with cardiovascular disease. Conversely, false-negative diagnosis is when the physician misdiagnoses a person with cardiovascular disease as healthy. Furthermore, the presence of outliers from the underlying databases decreases the accuracy of prediction models for reliable knowledge discovery. Due to the negative impacts, there remains a need to develop a new semi-supervised one-class classification algorithm for classifying conditional anomalous instances and improving multiple linear and logistic regression prediction models' accuracy.

2. RELATED LITERATURE

2.1 Data Mining

Data mining is used by machine learning as the information source to extract knowledge from a large amount of data (Witten & Frank, 2005). It is also an essential step of Knowledge Discovery in Databases that applies data analysis and algorithms that produce model over the data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Silwattananusarn & Tuamsuk, 2012).

Moreover, data mining tasks involve the generation of descriptive and predictive models (Fadzilah Siraj & Mansour Ali Abdoulha, 2011). The descriptive methods identify relationships in data, while predictive methods predict values, as shown in the literature map of the study (Figure 1).

The descriptive methods used in data mining are an association, clustering, summarization, and sequence analysis. Association finds relationships between multiple variables in a dataset while clustering partitions data points into sub-classes (Sagar et al., 2017). Clustering groups similar data to one another, where class labels of the data are not available (Patel & Mehta, 2011). Summarization extracts and characterizes the contents of the database (Dunham, 2002). Sequence analysis finds a relationship between data and periods (Rezig et al., 2018).

The predictive methods in data mining are classification, prediction, time series, and regression (Dunham, 2002; Fadzilah Siraj & Mansour Ali Abdoulha, 2011). Classification generates a model from a training subset that distinguishes classes and uses the model to predict unseen data, while prediction generates models for continuous and dichotomous response variables (Han, Pei & Kamber, 2006). Prediction determines the likelihood or outcome of the data (Fadzilah Siraj & Mansour Ali Abdoulha, 2011). Time series extracts information defined by the time they were recorded (Yan, Ulanova, Ouyang, & Xu, 2014). Regression predicts real-valued data items by some known type of function (Dunham, 2002). The linear & multiple regression, logistic, lasso, and ridge are the types of regression techniques to make predictions (Geetha, 2018).

2.2 Classification Algorithms

The classification algorithms provide models or classifiers that identify a set of categories (Khan & Madden, 2014). These classification models can be formed using an If-Then rule, decision tree, or neural network (Han, Pei & Kamber, 2006). Some of the classification algorithms are ID3, C4.5, Bayesian Network, K-Nearest Neighbor (KNN), and SVM (Soofi & Awan, 2017).

Classification algorithms will use a training subset to build a model or classifier. The model will learn by analyzing the training subset made of attributes with instances and their associated class labels. The class label attribute is discrete, unordered, and categorical. The testing subset will be used to evaluate the accuracy of the model (Han, Pei & Kamber, 2006).

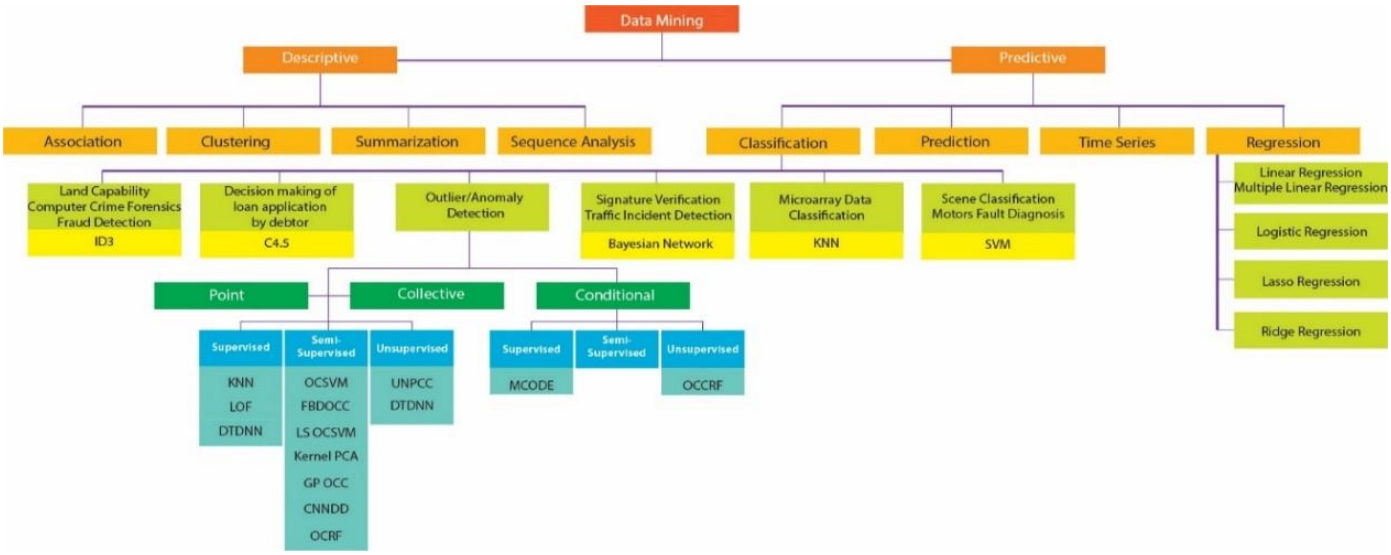

Figure 1: Literature Map

The classification algorithms are evaluated according to their classifier's accuracy, speed, robustness, scalability, and interpretability. The classifier's accuracy is evaluated to test its ability to classify unseen instances correctly. The

confusion matrix is used to evaluate for binary and multi-class classification algorithms (Tharwat, 2018).

Some applications of classification algorithms are land capability classification, computer crime forensics, fraud detection, decision making of loan application by debtor, signature verification, traffic incident detection, micro array data classification, scene classification, motors fault diagnosis (Soofi & Awan, 2017), analysis of road safety risk factor dependencies (Kwon, Rhee, & Yoon, 2015), disease and spam detection (Tong, Feng, & Li, 2018), movie review (Chaovalit & Zhou, 2005), text classification (Kowsari et al., 2019), economic forecasting (Bafandeh & Bolandraftar, 2013), fault diagnosis (Zhao, Yang, Lu, & Wang, 2015), cancer genomics (Huang et al., 2018), speech recognition (Ganapathiraju, Hamaker, & Picone, 2004), galaxy morphologies (Freed & Lee, 2013), hydrology (Raghavendra. N & Deka, 2014), bioinformatics (Wang et al., 2017), lung classification (Porkodi & Karuppusamy, 2019), bioinformatics (Vignesh et al., 2019), mushroom classification (Ottom, Alawad & Nahar, 2019), outlier or anomaly detection (Aggarwal & Heights, 2016) and many more.

2.3 One-Class

The single-class classification was introduced decades ago in the context of labeling only one class of crop, which is wheat in the Large Area Crop Inventory Experiment using Bayes classifier (T. C. Minter, 1975). The classifier learned from the labeled training samples of the class of interest. The classifier was able to minimize the need for ground-truth samples of other classes. The one-class classification originated from the research work of Moya et al. (1993), and similar studies were published, including novelty detection (Bishop, 1994), concept learning in the absence of counter-examples (Japkowica, 1999), and positive-only learning (Madden & Munroe, 2005).

The natural method of the one-class approach is distance-based (Eq. 1). The learned model will measure the unknown instance (x), and if it is smaller than the learned threshold, it will be classified as a normal class. Otherwise, it will be classified as another class (Camarinha-Matos, 2005).

$$\text{Class (x)} = \begin{cases} \text{target, if Measurement (x)} \leq \text{threshold;} \\ \text{Non-target, otherwise.} \end{cases} \quad (1)$$

2.4 Anomaly

Anomaly is considered as a special kind of outlier that is of interest with an analyst - discovering anomalous data points and instances or events that significantly deviate from the normal data point or events (Aggarwal & Heights, 2016).

There are three (3) types of anomalies, namely: point, contextual, and collective (Goldstein & Uchida, 2016), as shown in Figure 2.

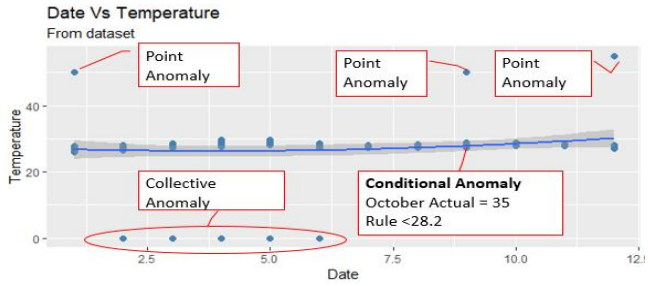

Figure 2. Type of Anomalies

The first type of anomaly is point anomaly, which occurs if an individual data instance deviates to the normal data points or events. The contextual or conditional anomaly occurs if the data instance is anomalous within a context or condition. This type of anomaly requires domain knowledge or requires a notion of context (Song, Wu, Jermaine, & Ranka, 2007). The last type of anomaly is called a collective anomaly. It contains the collection of data points or events that deviate from another group.

3. CONCEPTUAL FRAMEWORK

The concept used in the study is shown in Figure 3 below.

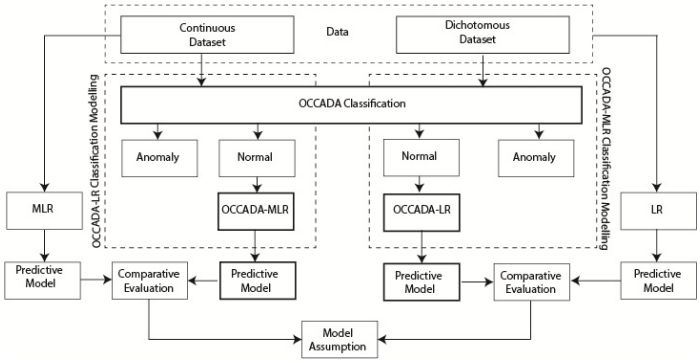

Figure 3: Conceptual Framework

The ground-truth and secondary datasets include response variables of continuous and dichotomous types. The training subsets were used to generate the OCCADA classification model. The OCCADA model used the testing subset to label the instances of the one-class datasets to either normal or anomalous class. All normal instances were used to generate the OCCADA-MLR predictive model for continuous response variables and the OCCADA-LR predictive model for dichotomous response variables. Then, a comparative evaluation of the predictive regression models was performed to determine the better model. The comparative evaluation between the OCCADA-MLR & classic MLR and OCCADA-LR & classic LR. Lastly, model assumptions for OCCADA-MLR and OCCADA-LR were performed.

4. MATERIALS

RStudio Version 1.2.1335, 2009-2019 RStudio, Inc. statistical tool was used to process the datasets. The computer system platform used was 64-bit Windows 10.

All processing was done on a Dell Inspiron 15 5000 series laptop running on Intel Core i7-7500 CPU @ 2.70GHz with 16.0GB of memory.

The experimental datasets used in the study were ground-truth climate change data and benchmark datasets for heart attack. The climate change data were collected from PAG-ASA XI and PNERC Station ID 537. The data collected from PAG-ASA used as indicator attributes temperature, relative humidity, rainfall, and daylight, which were extracted from 1990-2016. The response variable was the tide, which was collected from the PNERC and extracted from the database on August 12, 2019. The heart attack dataset was taken from the Kaggle repositories (Janosi, Steinbrunn, Pfisterer, Detrano, & Aha, 2019). The profiles of the datasets are shown in Tables 3 and 4.

Table 1. Characteristics of the datasets

Quantification Index	Climate Change	Heart Attack
Sample Size	180	261
No. of Attributes	8	7
Number of Classes	1	1
Area	Environment	Health

All instances in the datasets contain normal values extracted from the ground-truth collection and online repositories.

Table 4. Climate Change Dataset Structure

Attribute Name	Description	Type
year	Year (1990-2016)	Numeric
month	Month (1-12)	Numeric
season	Season (Dry, Rainy)	Condition
temp	Mean Temperature (Celsius)	Numeric
rh	Mean Relative Humidity (%)	Numeric
rainfall	Rainfall (mm)	Numeric
Daylight	Daylight (minutes)	Numeric
tide	Permanent Service for Mean Sea Level	Predicted attribute (Continuous)

The climate change dataset is a collection of two (2) datasets from PAG-ASA and PNERC. The records extracted from PAG-ASA were predictors *year*, *month*, *season*, *temp*, *rh*, *rainfall*, and *daylight*. The attribute tide was extracted from PNERC and used as a response variable in the study.

Table 5. Heart Attack Dataset Structure

Attribute Name	Description	Type
age	Age	Numeric
sex	Sex (0=Male, 1=Female)	Condition
trestbps	Resting blood pressure (mm)	Numeric
chol	Cholesterol (mg/dl)	Numeric
thalach	Maximum Heart Rate	Numeric
oldpeak	Depression induced by exercise relative to rest	Numeric
num	Diagnosis of heart disease (0=<50% diameter narrowing, 1=>50% diameter narrowing)	Predicted attribute (Dichotomous)

The heart attack dataset used five (5) indicator attributes, namely: *age*, *trestbps* for the resting blood pressure, *chol* for the serum cholesterol, *thalach* for the maximum heart rate achieved, and *oldpeak* for the depression induced by exercise relative to rest. The indicator attributes were selected because of their numeric form. Patient *sex* was used as the condition for the study, and the response variable was *num*, which contains a dichotomous type of values.

5. METHODS

5.1 Model Generation

There were six (6) models used to evaluate the classification and prediction accuracy of OCCADA, as shown in Table 6.

Table 6. List of Models

Model No.	Algorithm	Data Mining Technique	Dataset	Response Variable Type
1	OCCADA	Classification	Climate Change	Continuous
2	OCCADA	Classification	Heart Attack	Dichotomous
3	OCCADA-MLR	Regression	Climate Change	Continuous
4	MLR	Regression	Climate Change	Continuous
5	OCCADA-LR	Regression	Heart Attack	Dichotomous
6	LR	Regression	Heart Attack	Dichotomous

These models use two (2) datasets for classification and regression. The models one (1) and two (2) were used to evaluate the performance classification of OCCADA. Models three (3) and four (4) were utilized to compare the MLR using OCCADA from the classical MLR. The models' used climate change dataset because its response variable is continuous. The MLR technique uses a continuous numeric response variable for prediction. Models five (5) and six (6) were used to evaluate the prediction accuracy of LR using OCCADA over the classic LR. The heart attack dataset was used for the comparison because logistic regression predicts the dichotomous type of response variables.

5.2 Algorithms

5.2.1 Generate the OCCADA Model

The first algorithm was designed to generate the OCCADA models 1 and 2 for the climate change and heart attack datasets, as shown in the algorithm below.

Input:
D – given one-class multivariable dataset;
N – number of instances;
n_b – number of behavior attributes;
T- Threshold

Output: One-Class Conditional Anomaly Detection Model

1: createDataPartition D to n_b
2: **for** each D_{nb} do
3: createDataPartition D_{nb} into $D_{nbTrain}$ for training and D_{nbTest} for testing
4: **end for**
5: **for** each selected $D_{nbTrain}$
6: $D_{nbTrain}[N] > max$
7: max := $D_{nbTrain}[N]$
8: $D_{nbTrain}[N] < min$
9: min := $D_{nbTrain}[N]$
10: **end for**
11: return (max, min)
12: Obtain T
13: Combine max, min, and T

The *one-class multivariable dataset*, *user-specified behavior attribute*, and *threshold* were required as inputs. The holdout method was used to train and test the model.

5.2.2 Generating Balanced Algorithm

The second algorithm was designed to generate a balanced dataset to be able to evaluate the model's accuracy.

Input:
D_{nbtest} – testing datasets;
N – Number of instances;
n_b – Number of behavior attributes;
n_i- Number of indicator attributes

Output: Balanced Testing dataset

1: $D_{BalanceDataset}$ <- cbind($D_{nbtest1,}$ $D_{nbtestn}$)
2: **for** i in range (1,N) do
3: generate row instance number
4: assign $D_{BalanceDataset}$ $Class: =Normal
5: **end for**
6: $D_{anomaly} = N/2$
7: $D_{anomalybehavior} = D_{anomaly}/n_b$
8: $D_{anomalyindicator} = D_{anomalybehavior}/n_i$
9: **for** each $D_{anomaly}$
10: **for** each $D_{anomalybehavior}$
11: **for** each $D_{anomalyindicator}$
12: generate random instance number [$D_{anomalyindicator}$]
13: $D_{BalanceDataset}$[N,$D_{anomalyindicator}$]<-sample[<min:>max, $_{anomalyindicator}$]
14. $D_{BalanceDataset}$ [N, $D_{anomalyindicator}$]$Class = "Anomaly"
15: **end for**
16: generate random instance number [$D_{anomalybehavior}$]
17: $D_{BalanceDataset}$ [N, $D_{anomalybehavior}$] <- sample[<min:>max, $D_{anomalybehavior}$]
18: $D_{BalanceDataset}$ [N, $D_{anomalybehavior}$]$Class = "Anomaly"
19: **end for**
20: generate random instance number [$D_{anomaly}$]
21: $D_{BalanceDataset}$ [N, $D_{anomaly}$] <- sample[<min:>max, $D_{anomaly}$]
22: $D_{BalanceDataset}$ [N, $D_{anomaly}$] $Class = "Anomaly"
23: **end for**

The algorithm two (2) was developed to evaluate the models 1 and 2 by using a balanced dataset with 50% artificially injected anomalies. The purpose of having a balanced data set is to remove or minimize biased predictions that may mislead the accuracy of the model (Mishra, 2017). The algorithm requires the training dataset as input. A randomize instance number was generated, and normal instances were replaced with anomalies. The model was applied to the balanced testing dataset. The test data contains 40% of the original data, which holds normal class with no missing values and outlier-free. The model must be evaluated to test its accuracy using the test data injected with conditional anomalous values.

5.2.3 Model Classification and Evaluation

The evaluation takes the balanced testing dataset and the OCCADA model as input. The instances that conform to the model was labeled normal in the class attribute. Otherwise, instances that do not conform to the behavior was labeled anomaly.

Moreover, the confusion matrix was used to summarize the number of True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN) in the confusion matrix. The confusion matrix evaluates the classifiers correct, TP and TN, and incorrect, FP and FN, classification of instances. Thus, the improvement of OCCADA was evaluated using the confusion matrix. Furthermore, TPR, TNR, FPR, FNR, accuracy, and F1-score were measures derived from the confusion matrix.

5.2.4 Multiple Linear Regression with OCCADA

The application of OCCADA to MLR using the continuous response variables was processed using the algorithm below.

Input:
$D_{BalanceDataset} = D_{BalDat}$;
N = number of instances;
IV = Independent variables;
DV = Dependent variables;
lm = Linear Regression model

Output: OCCADA-MLR Prediction

1: Trim $D_{BalDat}$$Class <- "Anomaly"
2: createDataPartition D_{BalDat} into $D_{BalDatTrain}$ for training dataset and $D_{BalDatTest}$ for Testing dataset
3: Obtain IV and DV
4: fit <- lm(y ~ 0+ IV , data= $D_{BalDatTrain}$)
5: predict <- predict(fit, newdata = $D_{BalDatTest}$)
6: actuals_preds <- data.frame(cbind(actuals, predicteds))
7: error <- actuals_preds$actuals - actuals_preds$predicteds
8: AIC(fit), BIC(fit), RMSE(error), MAPE(error)

The two (2) datasets used in generating the models 3 and 4 were climate change dataset applied with OCCADA and climate change without OCCADA. These datasets were fitted using the *lm* multiple regression function of R. The output *fit* models were compared in terms of AIC, BIC, RMSE, MAE, and MAPE to determine the better model for MLR.
The *modelr* package of R provides functions such as AIC and BIC for computing the regression model performance metrics. The models with the lowest AIC and BIC score is preferred (Witten, Frank, & Hall, 2011).
Moreover, the Global Validation of Linear Model Assumptions (GVLMA) was implemented for testing the assumptions of the linear model (Peña & Slate, 2006).

5.2.5 Logistic Regression with OCCADA

The application of OCCADA was applied to LR using algorithm five (5), as shown below.

Input:
$D_{BalanceDataset}$ = D_{BalDat};
N = number of instances;
IV = Independent variables;
DV = Dependent variables;
glm= Generalized Linear Regression model

Output: OCCADA-LR Prediction

1: Trim $D_{BalDat}$$Class <- "Anomaly"
2: createDataPartition D_{BalDat} into $D_{BalDatTrain}$ for training dataset and $D_{BalDatTest}$ for Testing dataset
3: Obtain IV and DV
4: fit <- glm (y ~ IV , family = "binomial", data= $D_{BalDatTrain}$)
5: predict <- predict(fit, newdata = $D_{BalDatTest}$)

The algorithm five (5) was developed to extend the application of OCCADA to LR. It takes as input the balanced dataset using a dichotomous response variable. The dataset consists of normal values without conditional anomalies. Also, algorithm uses the generalized linear regression model package using the binomial family to predict the dichotomous response variables. The accuracy of the OCCADA-LR was measured using null deviance, residual deviance, AIC, and accuracy.

The generalized linear model, *glm*, function of R was used to generate the models for the dataset. The algorithm generates a model using the *glm* function using the heart attack dataset applied with OCCADA. The *glm* function includes a parameter family binomial that can process a dichotomous response variable. This function automatically generates the values for the null deviance, residual deviance, and AIC. These are measures that determine the ideal model for LR. Also, the prediction accuracy of the model was calculated by comparing the fitted and actual values.
One of the assumptions of logistic regression is that there should be no high intercorrelations or multicollinearity of indicator variables (Daoud, 2018). The multicollinearity assumptions for OCCADA-LR were performed using the Variance Inflation Factor (VIF) package of R. This package was used to validate the results of the OCCADA-LR model.

6. RESULTS AND DISCUSSION

6.1 Basic Statistics Results and Discussions

The basic statistics and correlation of the Climate Change and Heart Attack datasets are shown in Tables 7-9, respectively.

Table 7. Basic Statistics of the Climate Change Dataset

n=180	year	month	temp	rh	rainfall	daylight	tide
Min	1990	1.00	26.40	68.00	1.40	7800.00	6929
1st Qu.	1993	4.00	27.80	78.00	90.47	10739.00	7109
Median	2007	7.00	28.20	81.00	143.75	12270.00	7179
Mean	2005	6.55	28.23	80.18	146.80	12279.00	7166
3rd Qu.	2013	9.00	28.70	82.0	196.82	13539.00	7229
Max	2016	12.00	30.60	91.0	430.50	18540.00	7349

The total number of instances (n=180) were collected from January to December of 1990 to 2016. Some instances were removed because of the null values reported from the station.

Table 8. Correlation of the Climate Change dataset

	temp	rh	rainfall	daylight	tide
temp	1.00	-0.45	-0.18	0.56	0.24
rh.	-0.45	1.00	0.50	-0.61	0.10
rainfall	-0.18	0.50	1.00	-0.33	0.20
daylight	0.56	-0.61	-0.33	1.00	-0.09
tide	0.24	0.10	0.20	-0.09	1.00

Similarly, the basic statistics results for the heart attack dataset was computed and the results are shown in Table 9-10.

Table 9. Basic Statistics of the Heart Attack Dataset

n=261	age	sex	trestbps	chol	thalach	oldpeak
Min	28.00	0.00	92.00	85.00	82.00	0.00
1st Qu.	42.00	0.00	120.00	208.00	122.00	0.00
Median	49.00	1.00	130.00	242.00	140.00	0.00
Mean	47.77	0.73	132.60	248.80	139.20	0.61
3rd Qu.	54.00	1.00	140.00	280.00	155.00	1.00
Max	65.00	1.00	200.00	603.00	190.00	5.00

The result of the correlation of the attributes is shown in Table 10.

	Model 1 OCCADA		**Model 2** OCCADA	
	Climate Change (Continuous)		Heart Attack (Dichotomous)	
	Actual Positive	Actual Negative	Actual Positive	Actual Negative
Predicted Positive	33	4	26	3
Predicted Negative	3	32	2	27

Table 10. Correlation of the Heart Attack Dataset

	age	**trestbps**	**chol**	**thalach**	**oldpeak**
age	1.00	0.26	0.09	-0.46	0.21
trestbps	0.26	1.00	0.12	-0.22	0.23
chol	0.09	0.12	1.00	-0.14	0.11
thalach	-0.46	-0.22	-0.14	1.00	-0.33
oldpeak	0.21	0.22	0.11	-0.33	1.00

The correlation shows positive and negative linear correlations between the attributes.

6.2 OCCADA Model

The first algorithm generated a model from the behavior and indicator attributes of climate change and heart attack datasets. The rules are shown in Tables 11 and 12 for climate change and heart attack, respectively.

Table 11. Climate Change Dataset

	temp	**rh**	**rainfall**	**daylight**	**tide**
A. Behavior 1 - Dry Season (January-May and December)					
Minimum	26.40	68.00	1.40	7800.00	6929
Maximum	30.60	89.00	430.5	18069.00	7319
Threshold	0.1°C▲				
B. Behavior 2 - Rainy Season (June to November)					
Min	27.40	72.00	27.70	8556.00	6959
Max	29.60	91.00	391.8	18540.00	7349
Threshold	0.1°C ▲				

The model includes the behavior seasons dry and rainy. The 0.1 °C was the result of an input threshold of the user that was based on the study of PAG-ASA. The same process was used for the generation of the model using the heart attack dataset as shown in Table 12.

Table 12. Heart Attack Dataset

	age	**tresbps**	**chol**	**thalac**	**oldpeak**
A. Behavior 1 - Female					
Minimum	31.00	100.00	160.00	90.00	0.00
Maximum	62.00	160.00	394.00	185.00	2.00
Threshold					
B. Behavior 2 - Male					
Minimum	32.00	98.00	85.00	98.00	0.00
Maximum	62.00	190.00	412.00	185.00	1.50
Threshold					

The models of the heart attack dataset consist of the behaviors of male and female and indicator attributes. There is no threshold provided for this dataset.

6.3 Model Classification and Evaluation

The result of the model classification and evaluation of the two (2) models is shown in Table 13.

Table 13. Confusion Matrix for Climate Change and Heart Attack Datasets

The first and second models generated higher TP and TN. Thus, lower values for FP and FN using the OCCADA model. Also, additional measures were used to evaluate the OCCADA model as shown in Table 14.

Table 14. Other Measures of the OCCADA using the Climate Change (Model 1) and Heart Attack (Model 2) Datasets

Measure	**OCCADA Model 1**	**OCCADA Model 2**
Sensitivity or True Positive Rate (TPR)	0.92	0.93
Specificity or True Negative Rate (TNR)	0.89	0.90
False Positive Rate (FPR)	0.11	0.10
False Negative Rate (FNR)	0.08	0.07
Accuracy	**0.90**	**0.91**
F1 Score	0.90	0.91

Both models generated high values in TPR, TNR, accuracy, and F1. Consequently, small values of the FPR and FNR reduces the error in the misclassification of positive or normal values.

6.5 MLR with OCCADA

The experimental result between the classic MLR and OCCADA-MLR is shown in Table 15.

Table 1. Multiple Regression Models Evaluation using Climate Change Dataset

Measure	**OCCADA-MLR Model 3**	**MLR Model 4**
AIC (Akaike Information Criterion)	139.00	694.71
BIC (Bayesian Information Criterion)	148.92	747.08
RMSE (Root Mean Square)	9.79	21.89
MAE (Mean Absolute Error)	6.57	17.37
MAPE (Mean Absolute Percentage Error)	0.09	0.24

The AIC, BIC, RMSE, MAE, and MAPE values must have low values to determine the better prediction model for MLR according to the study of Maindonald and Braun (2010). Based on the results, model 3 generated lower values for AIC, BIC, RMSE, MAE, and MAPE. Model 3 exhibited improvement over the classic MLR with percentage point changes of 79.99% for AIC; 80.07% for BIC; 55.28% for RMSE; 62.18% for MAE; and 62.50% for MAPE.

6.5.1 MLR Assumptions

The result of the gvlma() package is shown in Table 16.

Table 16. Assessment of the gvlma() Package of the MLR with OCCADA Model

Criteria	p-value	Decision
Global Stat	0.98	Assumptions acceptable
Skewness	0.95	Assumptions acceptable
Kurtosis	0.75	Assumptions acceptable
Link Function	1.00	Assumptions acceptable
Heteroscedasticity	0.65	Assumptions acceptable

The global stat's p-value > 0.05 indicates that there is a linear relationship between the variables. The results show that the OCCADA-MLR model did not violate the linear assumption. The skewness' and kurtosis' p-value > 0.05 determined the normal distribution of the data. The link function's p-value > 0.05 indicates a numeric dependent variable. The heteroscedasticity p-value > 0.05 indicates that the model is better in predicting for certain ranges.

6.6 Logistic Regression with OCCADA

Both models were executed several times and the mean values of the multiple runs are shown in Table 17.

Table 17. Results of LR Models

Measure	Model 5 (OCCADA-LR)	Model 6 (Classic LR)	Ideal Criterion
Null Deviance	80.40	224.58	Low
Residual deviance	10.61	129.61	Low
AIC	24.61	143.62	Low
Accuracy	0.95	0.82	High

The model with the lower values on null deviance, residual deviance, and AIC and high value on accuracy are considered as a better LR model (Aquila & Community, 2018; Sapra, 2013). Based on the results, OCCADA's Model 5 generated a better prediction model compared to Model 6. It showed that OCCADA-LR generated a percentage change on null deviance by 64.20%; residual deviance by 91.81%; AIC by 82.87%; and the accuracy increase percentage of 12.83%.

6.6.1 LR Assumptions

The results of vif() for the OCCADA-LR model are shown in Table 18.

Table 18. Results of vif () package of the LR with OCCADA Model

	Predictor Variables					
	x1	x2	x3	x4	x5	x6
VIF	1.48	1.00	1.17	1.61	2.89	2.91

A VIF value of 1 means predictor variables is not correlated. Also, the VIF value of $1 < VIF \leq 5$ means moderately correlated and VIF > 5 indicates a high correlation. Based on the results, there are no predictor variables values exceeding 5. Therefore, the OCCADA-LR model did not violate the multicollinearity regression assumption.

7. SUMMARY

The development of a new semi-supervised classification algorithm for one-class conditional anomaly detection was effective in classifying conditional anomalous instances and improves the predictive accuracy of MLR & LR on continuous & dichotomous response variables.

The hold-out method was effective in training and testing the classification and predictive regression models. The OCCADA model from the behavior attributes is a significant factor that increases the classification accuracy of the model. The higher values for sensitivity, specificity, accuracy, & F1 score lead and low false positive & false negative values leads to a model capable of attaining better classification performance.

Furthermore, this study was able to show that applying OCCADA to MLR and LR enables predictive models to generate better performance. The lower values for AIC, BIC, RMSE, MAE, and MAPE leads to a better MLR predictive model. Also, lower values of the null deviance, residual deviance, and AIC but higher in the accuracy measure will also lead to the preferred LR model.

Moreover, the R's gvlma()and vif() packages were effective in checking for the basic assumptions of the MLR and LR models. Overall, the experiments conducted showed that the OCCADA model was able to reduce the number of false-positive and false-negative classification errors. Also, comparative experiments conducted showed that MLR and LR, when applied with OCCADA, were able to improve predictive accuracy.

8. CONCLUSIONS

The use of OCCADA as classifier, when used on data with continuous response variable (Climate Change data), produced accurate results as shown by a sensitivity or TPR of 92%; a specificity or TNR of 89%; a False Positive Rate (FPR) of 11%; a False Negative Rate (FNR) of 8%; accuracy rates of 90%; and F1 scores of 90%.

The use of OCCADA as classifier, when used on data with dichotomous response variable (heart attack data), produced accurate results as shown by a sensitivity or TPR of 93%; a specificity or TNR of 90%; a False Positive Rate (FPR) of 10%; a False Negative Rate (FNR) of 7%; accuracy rates of 91%; and F1 scores of 91%.

The application of OCCADA in the MLR model on continuous response type of data resulted in improvement over the classical MLR, with improvements in the: AIC of 79.99%; BIC of 80.07%; RMSE of 55.28%; MAE of 62.18; and MAPE of 62.50%.

The Logistic Regression, when applied with OCCADA, exhibited improvement over the classical LR with the percentage points change on dichotomous response type of data using the following criteria: Better Null Deviance by 64.20%; Better Residual Deviance by 91.81%; Better AIC by 82.87%; and Increased Accuracy by 12.83%.

The conclusions above show that OCCADA is effective as a classification algorithm for instances with conditional anomalies. When applied towards prediction, the use of OCCADA in MLR was able to improve on the accuracy of the classical MLR when dealing with data with continuous response variables such as for climate change data. At the same time, it was also able to improve on the accuracy of the classical LR when dealing with data with dichotomous response variables such as for heart attack dataset. Thus, the objectives of the study have been achieved.

REFERENCES

[1] C. C. Aggarwal and Y. Heights. ***Outlier Analysis Second Edition***. Published by Springer, Cham, 2016, ch. 1, pp. 3-4.

[2] S. Bafandeh, and M. Bolandraftar. **Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background**, on Int. Journal of Engineering Research and Applications, Vol. 3, no. 5, pp. 605–610, 2013.

[3] C. M. Bishop. **Novelty detection and neural network validation**, *IEEE Proceedings - Vision, Image, and Signal Processing*, Vol. 141, No. 4, pp. 217, 1994. https://doi.org/10.1049/ip-vis:19941330

[4] P. Chaovalit and L. Zhou. **Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches**, in *Proc. of the 38th Annual Hawaii International Conference on System Sciences*, USA, 2005, pp. 112c-112c.

[5] J. I Daoud. **Multicollinearity and Regression Analysis**. *Journal of Physics: Conference Series*, 2018, Vol. 949, No. 1.

[6] M. H. Dunham. **Data Mining Introductory and Advanced Topics**, Pearson, 2002, ch. 1, pp. 5-9.

[7] F. Siraj and M. A. Abdoulha. **Mining enrolment data using predictive and descriptive approaches**. Knowledge-Oriented Applications in Data Mining, 2011, pp. 53–72. https://doi.org/10.5772/14210

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. **From data mining to knowledge discovery in databases.** *AI Magazine*, Vol. 17, No. 3, pp. 37–53, 1996.

[9] M. Freed and J. Lee. **Application of Support Vector Machines to the Classification of Galaxy Morphologies**, in *Proc. 2013 International Conference on Computational and Information Sciences, IEEE*, 2013, pp. 322-325.

[10] A. Ganapathiraju, J. Hamaker, and J. Picone. **Applications of Support Vector Machines to Speech Recognition.** *IEEE Transactions on Signal Processing*, Vol. 52, No. 8, pp. 2348–2355, August 2004. https://doi.org/10.1109/TSP.2004.831018

[11] M. Geetha, E. Shanti, and S.S Raman. **A Survey and Analysis on Regression Data Mining Techniques in Agriculture**. *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 8, pp. 341–347, January 2018.

[12] M. Goldstein and S. Uchida. **A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data**. *PLoS ONE*, Vol. 11, No. 4, April 2016.

[13] J. Han, J. Pei, and M. Kamber. ***Data mining : concepts and techniques***, Morgan Kaufman-Elsevier, 2006, ch. 1, pp. 15-20.

[14] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang and W. Xu. **Applications of Support Vector Machine (SVM) Learning in Cancer Genomics.** *Cancer Genomics & Proteomics*, Vol. 15, No. 1, pp. 41–51, 2018. https://doi.org/10.21873/cgp.20063

[15] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, and D. Aha. **Heart Disease UCI | Kaggle**, 2019.

[16] N. Japkowica. **Concept-Learning in the Absence of Counter-Examples: An Auto Association-Based Approach to Classification**, Ph.D. dissertation, Gradiate Program in Computer Science, The State University of New Jersey, 1999.

[17] Jiang, F., & Chen, Y.-M. **Outlier detection based on granular computing and rough set theory**. *Applied Intelligence*, Vol. 42, No. 2, March 2015.

[18] S. S. Khan and M. G. Madden. **One-class classification: Taxonomy of study and review of techniques.** *Knowledge Engineering Review*, Vol. 29, No. 3, pp. 345–374, January 2014.

[19] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown, **Text Classification Algorithms: A Survey**. *Information*, Vol. 10, No. 150, pp. 1-68, June 2019. https://doi.org/10.3390/info10040150

[20] O. H. Kwon, W. Rhee, and Y. Yoon, **Application of classification algorithms for analysis of road safety risk factor dependencies**. *Accident Analysis & Prevention*, Vol. 75, pp. 1–15, February 2015. https://doi.org/10.1016/j.aap.2014.11.005

[21] D. T. Larose and C. D. Larose. ***Data mining and Predictive Analytics,*** 2nd Ed., John Wiley & Sons, 2015, ch. 2.

[22] M. G. Madden and D. T.Munroe. **Multi-Class and Single-Class Classification Approaches to Vehicle Model Recognition from Images**, in Proc. of *AICS-05: Irish Conference on Artificial Intelligence and Cognitive Science*, Portstewart, September 2005.

[23] S. Mishra. **Handling Imbalanced Data: SMOTE vs. Random Undersampling**. *International Research Journal of Engineering and Technology*, Vol. 4, No. 8,

pp. 317–320, August 2017.

[24] V. R. Patel and R. G. Mehta. **Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm**. *International Journal of Computer Science Issues*, Vol. 8, No. 2, September 2011.

[25] E. A. Peña and E. H Slate. **Global Validation of Linear Model Assumptions.** *Journal of the American Statistical Association*, Vol. 101, No. 473, pp. 341, March 2006.

[26] S. N. Raghavendra, P. C. Deka. **Support vector machine applications in the field of hydrology: A review**. *Applied Soft Computing*, Vol. 19, pp. 372–386, June 2014. https://doi.org/10.1016/j.asoc.2014.02.002

[27] Rezig, S., Achour, Z., & Rezg, N. **Using Data Mining Methods for Predicting Sequential Maintenance Activities**. *Applied Sciences*, Vol. 8, No. 2184, November 2018.

[28] Sagar, P., Prinima, & Indu. **Analysis of Prediction Techniques based on Classification and Regression.** *International Journal of Computer Applications*, Vol. 163, No. 7, pp. 47–51, April 2017.

[29] S. K. Sapra. **Generalized Additive Models in Business and Economics**. *International Journal of Advanced Statistics and Probability*, Vol 1, No. 3, pp. 64–81, 2013.

[30] Y. Shi and L. Zhang. **COID: A cluster-outlier iterative detection approach to multi-dimensional data analysis**. *Knowledge and Information Systems*, Vol. 28, pp. 709-733, 2011. https://doi.org/10.1007/s10115-010-0323-y

[31] T. Silwattananusarn and K. Tuamsuk. **Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012**. *International Journal of Data Mining & Knowledge Management Process*, Vol. 2, No. 5, pp. 13–24, September 2012.

[32] Song, X., Wu, M., Jermaine, C., & Ranka, S. **Conditional Anomaly Detection**. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 5, pp. 631–645, March 2007. https://doi.org/10.1109/TKDE.2007.1009

[33] A. A. Soofi and A. Awan. **Classification Techniques in Machine Learning : Applications and Issues**. *Journal of Basic & Applied Sciences*, Vol. 13, pp. 459–465, August 2017.

[34] T. C. Minter. **Single-Class Classification**. *In Laboratory for Applications of Remote Sensing*, Vol. 75, pp. 12–15, 1975.

[35] A. Tharwat. **Classification assessment methods**. *Applied Computing and Informatics*, August 2018.

[36] D. Wang, S. Fong, R. K. Wong, S. Mohammed, J. Fiaidhi and K. K. L. Wong. **Robust high-dimensional bioinformatics data streams mining by ODR-ioVFDT**. *Scientific Reports*, Vol. 7, No. 43167, February 2017.

[37] I. H. Witten and E. Frank. ***Data Mining: Practical Machine Learning Tools and Techniques,*** *2nd ed.*, Morgan Kaufmann, 2005, ch. 1, pp. 4-9.

[38] I. H. Witten, E. Frank, and M. Hall. **Data Mining: Practical Machine Learning Tools and Techniques,** 3rd ed., Morgan Kaufmann, 2011, ch 5.

[39] T. Yan, L. Ulanova, Y. Ouyang, and F. Xu. **Data Mining in Time Series : Current Study and Future Trend**. *Journal of Computer Science*, Vol. 10, No. 12, pp. 2358–2359, 2014. https://doi.org/10.3844/jcssp.2014.2358.2359

[40] Z. Zhao, J. Yang, W. Lu, and X. Wang**. Application of local outlier factor method and back-propagation neural network for steel plates fault diagnosis,** *in Proc. IEEE: The 27th Chinese Control and Decision Conference*, 2015, pp. 2416–2421.

[41] X. Tong, Y. Feng, & J. Li. **Neyman-Pearson classification algorithms and NP receiver operating characteristics.** *Science Advances*, Vol. 4, No. 2, February 2018.

[42] V. Porkodi and S. A. Karuppusamy. **Classification of Chronic Obstructive Pulmonary Disease (COPD) Using Regression with Gabor Filtration and Random Forest Classification**. *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 5, pp. 2194-2198, 2019. https://doi.org/10.30534/ijatcse/2019/52852019

[43] U. Vignesh, G. Sivanageswara, B. M. Josephine, P. Nagesh. **Food Waste Protein Sequence Analysis using Clustering and Classification Techniques.** *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 5, pp. 2289-2298, 2019. https://doi.org/10.30534/ijatcse/2019/67852019

[44] M. A. Ottom, N. A. Alawad, K. M. O. Nahar. **Classification of Mushroom Fungi Using Machine Learning Techniques.** *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 5, pp. 2378-2385, 2019. https://doi.org/10.30534/ijatcse/2019/78852019