A New Multiclass Classification Method for Objects with Geometric Attributes Using Simple Linear Regression

Jeremias T. Lalis, *Member, IAENG*

***Abstract*—Data mining is very important for decision makers as it helps them in making proactive and knowledge driven decisions. Furthermore, its collection of techniques is used in diverse fields due to its applicability in real-world problems and these techniques can be grouped in to two, prediction and classification. Perceptron is considered as one of the earliest methods in classification that uses the concept of linear regression (LR) in learning the model. Support vector machine also uses the concept of LR but in a much complex manner. Some drawbacks in these methods are also discussed in this paper. This study presents a fast and easy to implement algorithm through simple LR in classifying multiclass objects with both linear and non-linear relationships among its classes. The algorithm exhibited a satisfactory performance based on the results of the experiments being conducted.**

***Index Terms*— Data mining, multiclass classification, simple linear regression, geometric properties**

I. Introduction

THE main goal of data mining is to uncover useful knowledge or pattern within the datasets that are rich with hidden information. Data mining has a collection of techniques that helps people in making proactive and knowledge driven decisions. These data mining techniques are being utilized by diverse fields for either prediction or classification purposes. Among these purposes, classification becomes one of the major researches due to its wide applications in real-world problems [1][3] such as natural resources recognition [13], spectral band selection [14], biomedical/biological modeling [15] etc.

Classification is a form of data analysis using supervised learning method. In this method, data containing observations are being analyzed to extract the model or function that can be used in determining the membership of the new observation into one of the predefined classes. Due to its various applications, it has been an active research topic in both statistics and machine learning fields [2]. Early works in statistical classification focused on discriminant function analysis (DFA), wherein the variables that can be used to discriminate between two or more classes are being identified. The main concept in this method is to use the identified predictor variables in the training set to construct discriminant functions, like linear functions, and determine the group membership of the unseen object. Modern statistical approaches focus on discovering more flexible class models like calculation of feature-space distance between the classes (e.g. K-Nearest Neighbor), estimation of features join distribution within each class (e.g. Bayesian) and classification tree analysis. In the other hand, machine learning field gives more attention in the construction of a more human-understandable classification expressions [2] like in the automatic generation of rule (e.g. decision tree), the use of conditional probabilities (e.g. Naïve Bayes), and even through linear and nonlinear regression (e.g. support vector machine) in creating more flexible models.

At the same time, linear regression is considered as a statistical approach and the most basic but widely used technique in prediction and modeling. In this technique, the best-fitting straight line, known as the regression line, are being calculated based on the presented feature space [12]. It aims to derive a linear function and reveal the linear relationship between the dependent variable x and independent variable y, denoted as

$$y = mx + b \tag{1}$$

where y is the criterion variable, m is the slope, x is the predictor variable, and b as y-intercept of the trendline.

In this case, it is called as simple linear regression since the value of the criterion variable is predicted based only on the value of predictor variable. Some other machine learning algorithms are also applying this concept in classifying linearly separable classes. Among these are perceptron and support vector machines (SVM). In these methods, the best-fitting line used to separate the two classes is referred as the hyperplane.

In data mining, classification and regression are both learning techniques that use the presented feature space in creating or learning the predictive models. However, these methods are used in different and specific purposes as it produce different values for output variables. Classification takes class labels as output, thus, it is used to find the class membership of the object. In the other hand, since regression takes continuous values as output, so it is used in estimating or predicting a response. However, there are some binary classifiers that adopted the concept of linear regression to classify objects but in a different and far more complex manner, these are perceptron and SVM.

Perceptron was invented by Frank Rosenblatt at the Cornell Aeronautical Laboratory in 1957 and considered as

Manuscript received on 20th February, 2016; revised 3rd March, 2016.

Jeremias T. Lalis is the current Research Director in La Salle University, Ozamiz City, Philippines. He is also an Assistant Professor in the College of Computer Studies in same university, e-mail: jeremias.lalis@gmail.com

the one of the earliest algorithms for linear classification [4]. It is a simple model of neuron that comprises of external input x that can be of any number, an internal input b, and one Boolean output value. Here, the suitable weight values w in the separating hyperplane $f(x)$ are computed so that the training examples can be correctly classified. The hyperplane is geometrically defined as

$$f(x) = \begin{cases} 1 & if \quad w \cdot x + b > 0 \\ 0 & otherwise \end{cases} \tag{2}$$

However, the major drawback in this algorithm is that the separating hyperplane is only guaranteed to be found if the learning set is linearly separable, otherwise, the training process will never stop. This made the perceptron less applicable to many pattern recognition problems having non-linear relationship between its attributes and class.

Like perceptron, support vector machine (SVM) is a linear, binary and hyperplane-based classifier. However, unlike perceptron, it is backed with solid theoretical grounding [5]. The goal in this algorithm is to find an optimal hyperplane, $w.x + b = 0$, that separates the two classes with the largest margin. It means that this hyperplane has the largest minimum distance to the training set. The hyperplane can be formally defined as

$$f(x) = sign\,(w^T \cdot x + b) \tag{3}$$

where w is the weight vector and b as the bias which can be computed by solving the constrained quadratic optimization problem based on the training data point. The final decision can then be derived and defined as

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i (x_i \cdot x) + b\right) \tag{4}$$

wherein, this function depends on a non-zero support vectors α_i which are often a small fraction of the original dataset.

SVM has been used already in wide variety of classification problems since its introduction in 1990s. However, some studies show some drawbacks of SVM. The limitation of SVM in terms of speed and size in both training and testing in large dataset was discussed in [16]. It was also pointed out that the optimal design for multiclass SVM classifiers requires further research.

Thus, in this study, the researcher aims to present a fast and easy to implement classification method for datasets with high-dimensional physical attributes based on simple linear regression. The researcher further aims to design the classifier to be accurate and precise even with the limited number of training dataset, around 20% of the dataset. This study also shows the applicability of simple linear regression in linear and nonlinear separable multiclass classification problems. Four (4) standard datasets from the UCI machine learning repository were used to measure and evaluate the performance of the proposed algorithm.

II. The Algorithm

Like the other classification methods, the proposed algorithm has two stages, namely the learning phase and classification phase.

A. Learning Phase

Any classifier should be trained first using the training set for it to be able to learn the predictive model and identify the correct class membership of the newly presented object. Fig 1 shows the block diagram of the training procedure of the proposed new classifier as presented in [17].

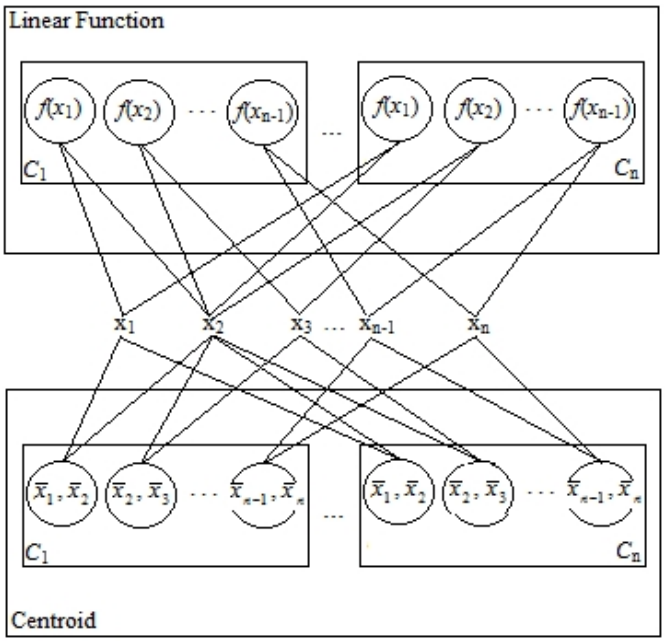

Fig 1. Block diagram of the proposed training procedure of the new classifier.

The training procedure comprises of two (2) simple steps:

Step 1: Find the linear relationship between the pairs of attributes in each class j based on the given training dataset with n number of x attributes and k tuples,

$$\begin{aligned} f_j(x_1) &= \alpha_1 \cdot x_1 + \beta_1 \\ f_j(x_2) &= \alpha_2 \cdot x_2 + \beta_2 \\ &\ldots \\ f_j(x_{n-1}) &= \alpha_{n-1} \cdot x_{n-1} + \beta_{n-1} \end{aligned} \tag{5}$$

where $f_j(x_i)$ is the linear function between attributes x_i and x_{i+1}, α_i is the slope, and β_i is the offset.

The slope α_i in $f_j(x_i)$ can be computed as:

$$\alpha_i = \frac{k\sum (x_i x_{i+1}) - \sum x_i \sum x_{i+1}}{k\sum x_i^2 - \left(\sum x_i\right)^2} \tag{6}$$

while the offset β_i in $f_j(x_i)$ is computed as:

$$\beta_i = \frac{\sum x_{i+1} - \alpha_i \sum x_i}{k} \tag{7}$$

The resulting values of α and β between the paired attributes in each class will then be used as internal inputs to

calculate the output value during the classification phase.

Figs 2, 3 and 4 illustrate the scatter plot of each pair of attributes as well as its corresponding regression line for each class using the Iris flower dataset.

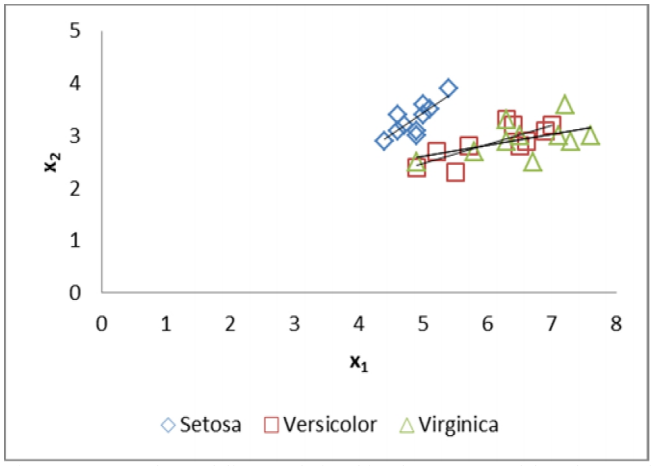

Fig 2. Scatter plot and linear relationships between sepal length x_1 and sepal width x_2 of the three classes.

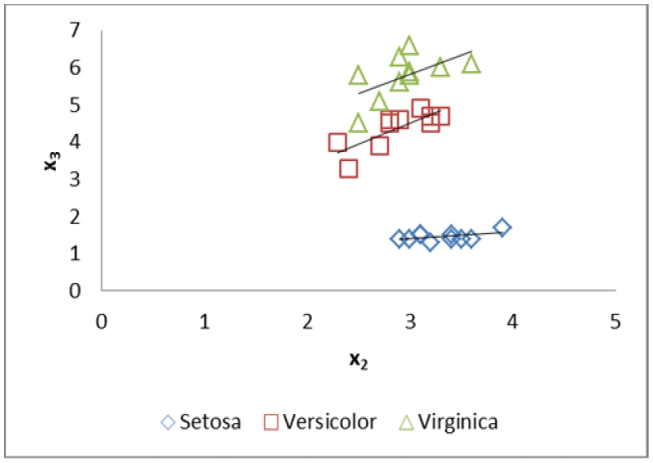

Fig 3. Scatter plot and linear relationships between sepal width x_2 and petal length x_3 of the three classes.

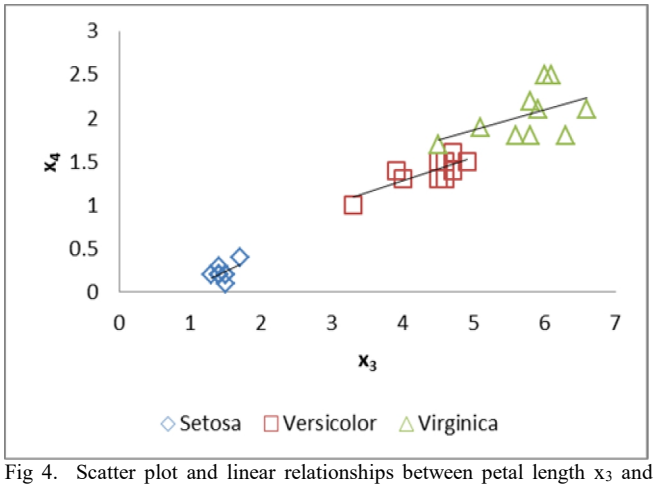

Fig 4. Scatter plot and linear relationships between petal length x_3 and petal width x_4 of the three classes.

Step 2: For each class *j*, compute the centroid *C* of the paired variables x_i and x_{i+1}, denoted as $C_j(\bar{x}_i, \bar{x}_{i+1})$:

$$\bar{x}_i = \frac{\sum x_i}{n}, \bar{x}_{i+1} = \frac{\sum x_{i+1}}{n} \quad (8)$$

Figs 5, 6 and 7 show the centroid of each pair of attributes. The learned model based on the sample data is shown in Table 1.

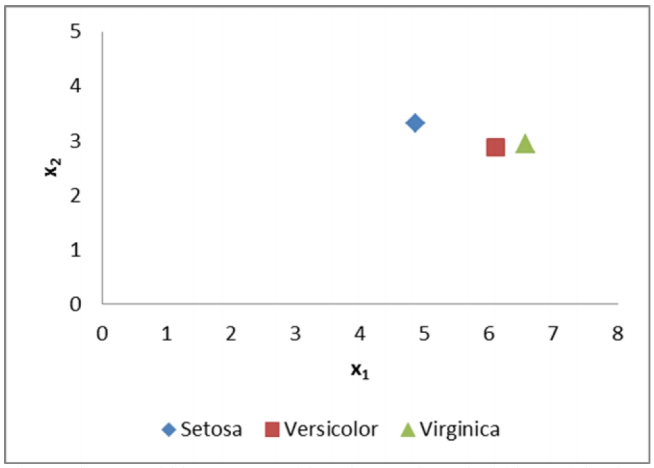

Fig 5. The centroid between sepal length x_1 and sepal width x_2 of the three classes.

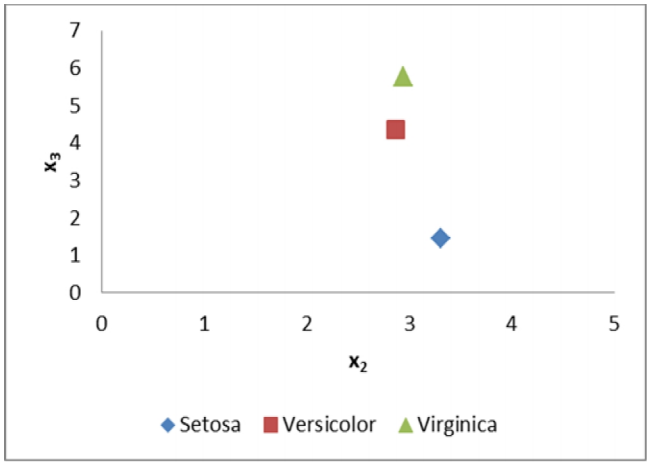

Fig 6. The centroid between sepal width x_2 and petal length x_3 of the three classes.

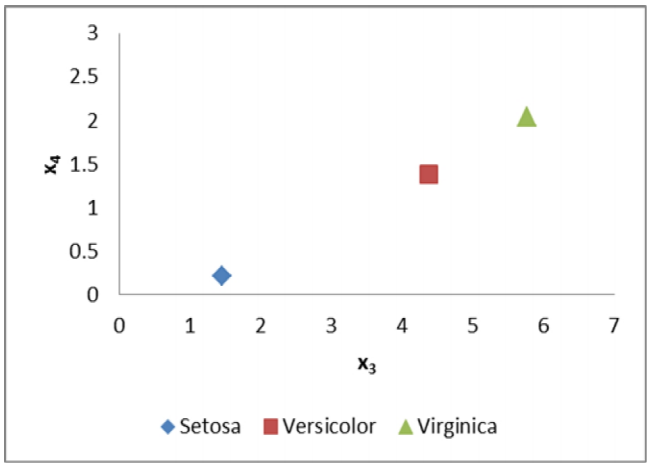

Fig 7. The centroid between petal length x_3 and petal width x_4 of the three classes.

B. Classification Phase

After the training process, the resulting model can now be used to classify the new object.

The following steps are used to determine the class membership of the input object and Table 2 illustrates the results of this process:

Step 1: The previously computed centroid C in every paired attribute in class j will serve as the first point (Point A) of the corresponding triangles, in the form of (internal input $\overline{x}_i$, internal input $\overline{x}_{i+1}$).

Step 2: Use the previously calculated linear functions $f_j(x_1), f_j(x_2), f_j(x_3) \ldots f_j(x_{n-1})$ and its corresponding input values $x_1, x_2, x_3, \ldots, x_{n-1}$ to find the second point (Point B) of the triangle for every paired attributes on its respective class j. The resulting xy-coordinates would be on the form of (external input x_i, internal input $f(x_i)$).

Step 3: Get the third point (Point c) of the corresponding triangles in class j by pairing the input values, e.g. (external input x_1, external input x_2).

Step 4: Calculate the area of its corresponding triangles by using the three (3) points on each paired attributes in class j,

$$\Delta\text{Area}_i = \left| \begin{pmatrix} x_i(\overline{x}_{i+1} - x_{i+1}) + \\ \overline{x}_i(x_{i+1} - f(x_n)) + \\ x_i(f(x_n) - \overline{x}_{i+1}) \end{pmatrix} \Big/ 2 \right| \tag{9}$$

Step 5: Get the total of all the corresponding ΔArea of its paired attributes to get the distance of the input object from the feature vectors in every class j,

$$dist_j = \sum_{i=1}^{n-1} \Delta\text{Area}_i \tag{10}$$

where n is the number of attributes.

Step 6: The class that obtained the least distance will be declared as the winner or the class membership of the new object.

TABLE I
MODEL LEARNED DURING TRAINING PHASE

j	i	α_i	β_i	C_j (Point A)	
				x-axis	*y-axis*
	1	0.8298	0.7230	4.86	3.31
Setosa	*2*	0.1826	0.8457	3.31	1.45
	3	0.3810	-0.3324	1.45	0.22
	1	0.3655	0.6402	6.10	2.87
Versicolor	*2*	1.1441	1.0865	2.87	4.37
	3	0.2728	0.1880	4.37	1.38
	1	0.2151	1.5269	6.57	2.94
Virginica	*2*	1.0273	2.7496	2.94	5.77
	3	0.2320	0.7012	5.77	2.04

TABLE II
CLASSIFICATION RESULT WHEREIN CLASS SETOSA IS THE WINNER

j	i	C_j (Point A)		Point B		Point C		Δ_j	$\sum_j$
		x-axis	*y-axis*	x_i	$f_j(x_i)$	x_i	x_{i+1}		
Setosa	*1*	4.86	3.31	5.1	3.47	5.1	3.5	0.00	
	2	3.31	1.45	3.5	1.11	3.5	1.4	0.03	**0.04**
	3	1.45	0.22	1.4	-0.25	1.4	0.2	0.01	
Versicolor	*1*	6.10	2.87	5.1	1.69	5.1	3.5	0.91	
	2	2.87	4.37	3.5	6.09	3.5	1.4	1.48	2.92
	3	4.37	1.38	1.4	0.56	1.4	0.2	0.54	
Virginica	*1*	6.57	2.94	5.1	2.16	5.1	3.5	0.99	
	2	2.94	5.77	3.5	8.68	3.5	1.4	2.04	5.15
	3	5.77	2.04	1.4	1.17	1.4	0.2	2.13	

III. EXPERIMENTS

A. Dataset

Four public datasets from UCI Machine Learning Repository were considered to measure and validate the performance of the new classification algorithm. The datasets are Iris Flower [6], Wheat Seed Kernel [7], Breast Tissue [8], Breast Cancer Wisconsin (Diagnostic) [9], and One Hundred Plant Species Leaves [10]. The characteristics of every dataset that was used in the experiments are shown in Table III.

TABLE III
DATASET CHARACTERISTICS

Dataset	Testing Size	Testing Size	# of Classes	Dim
Iris Flower	10 per class	40 per class	3	4
Wheat Seed	14 per class	56 per class	3	7
Breast Tissue	4 for class 1 10 for class 2 3 for class 3 4 for class 4	17 for class 1 39 for class 2 11 for class 3 18 for class 4	4	4
Breast Cancer	71 for class 1 42 for class 2	286 for class 1 170 for class2	2	30
Leaves-Shape	3 per class	13 per class	25	64

B. Evaluation

The 5-fold cross-validation was used in each experiment to evaluate the performance of the new classification algorithm. This means that the training and testing phases were performed five times by partitioning the dataset into five mutually exclusive subsets or folds.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Equation (9) was used to calculate the accuracy of the new classifier. This was used to measure how far off the predicted value is from the actual known value.

TABLE IV
CONFUSION MATRIX

	Condition Positive	Condition Negative
Test Outcome Positive	True Positive (*TP*)	False Positive (*FP*)
Test Outcome Negative	True Negative (*TN*)	False Negative (*FN*)

The confusion matrix in Table IV was used to further measure the performance of the new classifier in producing the model in every experiment. Equations (10), (11) and (12) for precision, recall and F-score (*F*) were used to measure the exactness, completeness and retrieval performance of the new classifier; respectively:

$$precision = \frac{TP}{(TP + FP)} \tag{10}$$

$$recall = \frac{TP}{(TP + FN)} \tag{11}$$

$$F = \frac{precision * recall}{precision + recall} \tag{12}$$

IV. RESULTS AND DISCUSSION

The summary of the results of experiments using the four datasets is reported in Table V.

TABLE V
EXPERIMENTS RESULTS SUMMARY

Dataset	Mean Accuracy %	Mean Precision %	Mean Recall %	Mean F-Score %
Iris Flower	94.50	95.05	94.50	94.77
Wheat Seed	89.27	89.72	89.23	89.48
Breast Tissue	75.29	75.13	67.88	71.32
Breast Cancer	88.55	89.92	85.91	87.87
Leaves (Shape)	83.69	86.47	83.63	85.03

Based on the experiments being conducted and the results obtained from it, the new classification algorithm performs best with the Iris flower dataset among the other three datasets. This proves that the simple linear regression is also applicable in classifying not only linearly separable, but including nonlinearly separable classes; followed with the experiments results conducted with the breast cancer dataset at a mean precision of 89.92%. It is also worth to note that the dataset distribution for training and testing is slightly imbalanced, wherein, only 38% is coming from the malignant class and the rest is from the benign class. However, the classifier performs better in the experiments performed using the wheat seed dataset in terms of mean *F*-score, mean accuracy and mean recall at 89.48%, 89.27% and 89.23%, respectively; compared to the results with the breast cancer dataset.

Note that the experiments conducted using the leaves dataset produced an acceptable result in terms of precision at 86.47% despite of being highly-dimensional and limited number of training set, wherein, there are only three (3) tuples for each class. In this dataset, many of the sub species resemble close appearance with the other major species and many sub species resemble a radically different appearance with its major species [11]. These added to the difficulty of the classification problem. The results also show the robustness of the new classifier by using the shape-based dataset only during the training and classification stage.

However, experiments conducted using the breast tissue dataset produced the lowest score, especially in terms of completeness at 67.88%. This is probably due to the unbalanced distribution of samples among its classes.

In general, the proposed algorithm performs satisfactorily even with small number of training set at 20% of the total size on each dataset.

V. CONCLUSION

This paper has presented a fast and easy to implement method that can be used for multiclass classification problems for objects with geometric attributes. This also proves the applicability of simple linear regression in both linearly and nonlinearly separable classes even though it was originally designed for binary classification problem and with linearly separable classes only. Empirical results show the satisfactory performance of the proposed algorithm using the four standard and public datasets taken from UCI machine learning repository.

For the future work, several avenues for improvement can still be considered like using the nonlinear regression to cater those paired attributes with nonlinear relationship.

REFERENCES

[1] V. S. M. Tseng and C. Lee, "Cbs: A new classification method by using sequential patterns," in *Proc.2005 SIAM International Data Mining Conference*, CA, 2005, pp596-600.

[2] A. An, Classification methods. CA: Idea Group Inc, 2005, pp. 1-6.

[3] A. Arakeiyan, L. Nerisyan, A. Gevorgyan, and A. Boyajyan, "Geometric approach for gaussian-kernel bolstered error estimation for linear classification in computational biology," *International Journal "Information Theories and Applications"*, vol. 21, no. 2, pp 170-181, 2014.

[4] F. Rosenblatt, "The perceptron-a perceiving and recognizing automation," Cornell Aeronautical Laboratory, New York, Report 85-460-1, 1957.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp273-297, 1995.

[6] UCI Machine Learning Repository, "Iris data set," 1988. [Online]. Available: archive.ics.uci.edu/ml/datasets/Iris

[7] UCI Machine Learning Repository, "Seeds data set," 2012. [Online]. Available: archive.ics.uci.edu/ml/datasets/seeds

[8] UCI Machine Learning Repository, "Breast tissue data set," 2010. [Online]. Available: archive.ics.uci.edu/ml/datasets/Breast+Tissue

[9] UCI Machine Learning Repository, "Breast cancer wisconsin (diagnostic)," 1995. [Online]. Available: archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[10] UCI Machine Learning Repository, "One-hundred plant species leaves data set," 2012. [Online]. Available: archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set

[11] C. Mallah, "Probabilistic Classification from a K-Nearest-Neighbor Classifier," *Computational Research*, vol. 1, no. 1, pp. 1-9, 2013.

[12] Wikipedia, "Linear regression," 2015. [Online]. Available: http://en.wikipedia.org/wiki/Linear_regression

[13] M. Shahbaz, A. Guergachi, A. Noreen and M. Shahee, "A data mining approach to recognize objects in satellite images to predict natural resources," *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 229, 2013, pp 215-230.

[14] S. A. Medjahed, T. A. Saadi, A. Benyettou and M. Ouali, "Binary cuckoo search algorithm for band selection in hyperspectral image classification," *IAENG International Journal of Computer Science*, vol. 42, no. 3, pp183-191, 2015.

[15] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2014, WCECS 2014, 22-24 October, 2014, San Francisco, USA, pp809-812.
[16] C. J. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*. Netherlands: Kluwer Academic Publishers, 1998, pp1-43.
[17] J. T. Lalis, "X-attributes classifier (XAC): a new multiclass classification method by using simple linear regression and its geometrical properties," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2015, WCECS 2015, 21-23 October, 2015, San Francisco, USA, pp776-780.

Jeremias T. Lalis (M'14), became a Member (M) of IAENG in 2014. He was born in Manila, Philippines on April 20, 1984. He received the B.S. degree in computer science from La Salle University, Ozamiz City, Philippines in 2005. He received the Masters and Doctorate degree in Information Technology from Cebu Institute of Technology-University, Cebu City, Philippines in 2011 and 2015, respectively. Now, he is an Assistant Professor at the College of Computer Studies, La Salle University, Ozamiz City, Philippines. His research interests lie on the area of image processing, computer vision, machine learning and data mining.