Improving Credit Scoring Performance using Two-Stage Technique

By
Ibrahim Anas and **Sam Olu Olagunju**

Department of Mathematics, Adeyemi Federal University of Education, Ondo.
anaseen1988@gmail.com // **lagsam2016@gmail.com**

Abstract
In this study an unsupervised learning based on Self-Organizing Map (SOM) was used to specifically improve the discriminant capabilities of Classification And Regression Trees (CART) and Artificial Neural Networks (ANN) used to predict the credit risk of borrowers from Bank of Agriculture (BOA) Sokoto. In this work, a two-stage approach to building the credit scoring model was proposed using SOM and CART. Within the two-stage scheme, the knowledge (i.e., prototypes of clusters) found by SOM were considered as input to the subsequent classification model (i. e. CART). The results from BOA, Sokoto data indicate that the two-stage model improved the performances CART from 96.3% to 96.7%. This therefore suggests that the integration of SOM algorithm into CART made SOM+CART hybrid model to outperform the stand-alone CART model.

Keywords*: Credit Scoring, Self-Organizing Map (SOM), Classification and Regression Tree (CART), Creditworthy, Non-Creditworthy.*

Introduction

Durand (1941) as the founder of credit scoring modelling was the first to recognize that one should differentiate between good or bad loans by measurements of the applicants' characteristics. Credit scoring can be formally defined as a statistical (or quantitative) method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent (Suleiman et al, 2017). This helps to determine whether credit should be granted to a borrower. Credit Scoring evaluation is one of the most crucial processes in bank credit risk management decisions, including collecting, analyzing, and classifying different credit elements and variables to measure the credit decisions (Suleiman et al, 2021).

A Self-Organizing Map (SOM) invented by Kohonen (1982) is a type of ANN (Artificial Neural Network) where the neurons are set along a grid. SOMs are different from other ANNs in the sense that they use a neighbourhood function to preserve the topological properties of the input space. It provides a visual way to understand high-dimensional data into a low-dimensional output space (Ali et al, 2016). The principal objective of SOM is to transform a complex high-dimensional input space into a simpler low-dimensional (typically two-dimensional) discrete output space by preserving the relationships (i. e. topology) in the data, but not the actual distances.

Classification and Regression Tree (CART) is a term used to describe decision tree algorithms that are used for classification and regression learning task. **The Classification and Regression Tree**

methodology, also known as the CART, which was introduced by Leo et al (1984), is a nonparametric method that employs binary trees and classifies a dataset into a finite number of classes. Vesanto and Esa (2000) Suggested, two-stage procedure; first using SOM to produce the prototypes that are then clustered in the second stage, these are found to perform well when compared with direct clustering of the data and to reduce the computation time. Huysmans et al (2005) showed how a trained SOM can be used for clustering. It uses feedforward neural networks as classification learning algorithm; the result showed that, the integration of a SOM with a supervised classifier is feasible. Reza and Hermawan (2017) proposed a hybrid method using CART algorithm and Binary Particle Swarm Optimization, the proposed method accuracy is 78 % and 87.53 % respectively. In comparison to several popular algorithms, such as neural network, logistic regression and support vector machine, the proposed method showed an outstanding performance. Mohammad (2015) aimed to compare the predictive capability of Logistic Regression, CART and Random Forest in classifying good and bad applicant classification. His results indicate that CART performance seemed to be better at 73.5% accuracy. Subsequently, the principal idea behind credit risk management is to predict creditworthy applicant accurately, the CART which seemed to perform better than other classification models, can be improved by filtering out non representative samples from the data using Self Organizing Map as proposed by Ali et al (2016). Therefore, the present work suggests, SOM+CART hybrid model which may lead to an outcome that will improve classification performance accuracy. This research aimed at improving Credit Scoring classification using two-stage technique by introducing SOM to CART.

Methodology

(a) Self-Organizing Map (SOM)

A SOM invented by Kohonen (1982) is able to reduce the amount of data and simultaneously project the data nonlinearly onto a lower dimensional array (see Figure1). In each iteration of the training process, the reference vectors are updated in such a way that the best-matching neuron and its neighbours on the grid are dragged toward the input. As a result, the neurons are topologically ordered on the grid, where instances that have similar features in the input space will be projected to the neurons located close to each other in the grid space (Ali, et al, 2016).

Since SOM is able to reduce the amount of data and simultaneously project the data nonlinearly onto a lower dimensional array, the neurons are distributed on a regular grid of usually two dimensions. We therefore have to chose the nearest neighbourhood for grouping or mapping similar neurons in order to reduce the dimensions of the data. Since the best-matching neuron and its neighbours on the grid are dragged towards the input, the neurons are topologically ordered on the grid where instances that have similar features in the input space will be projected close to each other in the grid space, resulting in dimensionality reduction of the data. This happens to be the main function of SOM.

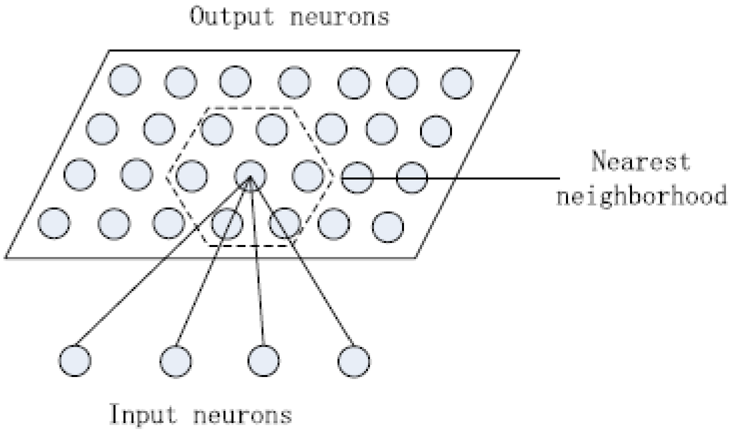

Figure1:
Example of self-organizing map composed of 4 input neurons and an output grid. (Ali, et al, 2016).

SOM Algorithm (**Model Training Process**) steps are as follows:

1. Initialize Neural Network weights
2. Randomly select an input
3. Select the winning neuron using Euclidean Distance:

$$d_j = \sqrt{\sum_{i=1}^{n}\left(x_i - w_{k,i}\right)^2}$$

where d_j represent the Euclidean distance for difference between each input each neuron, x represents the input,
w represents neuron weight, $k = 1, \ldots, m$ number of neurons,
and $i = 1, \ldots, n$ number of inputs.

4. Update neuron weight, using weight update formula:

$$\Delta w_{j,i} = \eta(t) * T_{j,I(x)}(t) * (x_i - w_{j,i})$$

where $\eta(t) = \eta_0 exp\left(-\frac{t}{\tau_\eta}\right)$ is the learning rate determining how quickly is the weight update, and $T_{j,I(x)}(t) = exp\left(-\frac{S^2_{j,i(x)}}{2\sigma(t)^2}\right)$ is the Topological Neighborhood, while $(x_i - w_{j,i})$ is the difference between the input and the weight.

5. Go back to 2 until done training.

(b) Classification and Regression Tree (CART)

Classification and Regression Trees is a classification method which uses historical data to construct so-called decision trees. Decision trees are then used to classify new data. In order to use CART, we need to know number of classes. Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts. CART asks only yes/no questions. CART algorithm will search for all possible variables and all possible values in order to find the best split, the question that splits the data into two parts with maximum homogeneity, this process is then repeated for each of the resulting data fragments (Kalamkas and Gulnar, 2013). The CART method employs binary trees and classifies a data into a finite number of classes, as such it is suitable for use in credit scoring where the default and non-default responses are contained in the data.

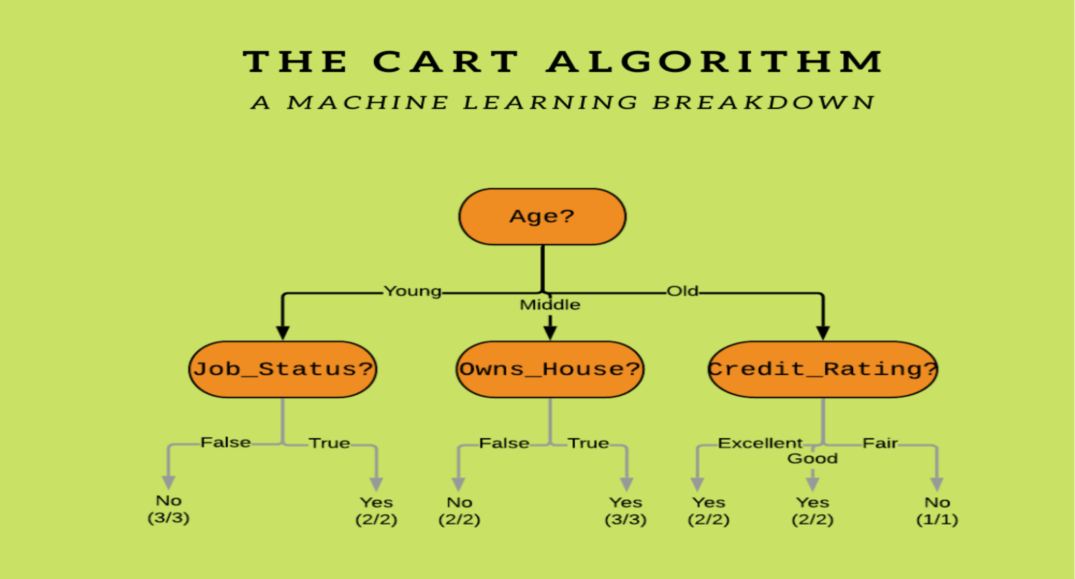

Figure2: Classification and Regression Tree Structure (iq.opengenus.org)

The procedure of fitting a classification tree to a data set is governed by three decisions:
(1) Splitting rule: According to which criterion should a group be split into two subgroups.
(2) Stopping rule: How to decide which subgroup is the terminal node. i. e. it should not be split any further.
(3) Classifying rule: How to assign a class (good loan/bad loan) to the terminal node. (Mohammad, 2015).

To obtain a pure leaf node for a separate/ binary classification tree, the impurity of the attribute needs to be measured using “Gini Index/ Gini Impurity” for each class or binary query.
This is given as: $G = 1 - (p(Y)^2 - p(N)^2)$
where *G* is the Gini Index, *p(Y)* is the probability of Yes response and *p(N)* is the probability of No.

(c) Evaluation Measure of Classification Models

In order to evaluate a binary decision task, we defined the following three performance metrics:

$Accuracy = \frac{tp + tn}{tp + fn + tn + fp}$ denoting the proportion of correct classifications out of the total samples.

$Sensitivity = \frac{tp}{tp + fn}$ denoting the fraction of true positives that are actually positive.

$Specificity = \frac{tn}{tn + fp}$ denoting the fraction of true negatives that are actually negative.

where tp refers to true positive, tn true negative,
fp is the false positive and fn refers to false negative.

Results and Discussions

The data used in this research is a secondary credit data set, extracted from the loan application forms of Agricultural and Rural Development Bank Sokoto, containing 300 cases. 164 applicants were considered as "Creditworthy" and the rest 136 were treated as "Non-Creditworthy". After data preparation, it was used for conducting the analysis using Self-Organizing Map (SOM) and Classification and Regression Tree (CART).

Table1: Description of the Dataset

S/No.	Variable	Type	Scale	Description
1.	Variable1	Input Variable	Scale	Applicant Age
2.	Variable2	Input Variable	Nominal	Applicant Gender/Sex
3.	Variable3	Input Variable	Nominal	Applicant Marital Status
4.	Variable4	Input Variable	Ordinal	Applicant Job
5.	Variable5	Input Variable	Nominal	Loan Purpose
6.	Variable6	Input Variable	Scale	Loan Credit Amount
7.	Variable7	Input Variable	Scale	Estimated Annual Salary of the Applicant
8.	Variable8	Input Variable	Nominal	Plan of the loan Repayment
9.	Variable9	Input Variable	Nominal	Type of Application
10.	Variable10	Input Variable	Nominal	Period of the Application
11.	Variable11	Output Variable	Nominal	Status of the Credit Applicant

(a) **Visualization of Self Organizing Map (SOM)**

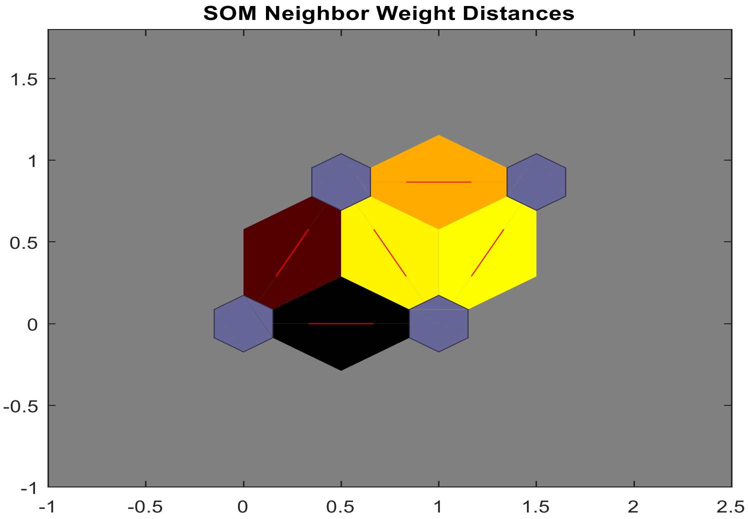

Figure3: SOM Neighbor Weight Distances

Figure3 indicates the distances between neighboring neurons.

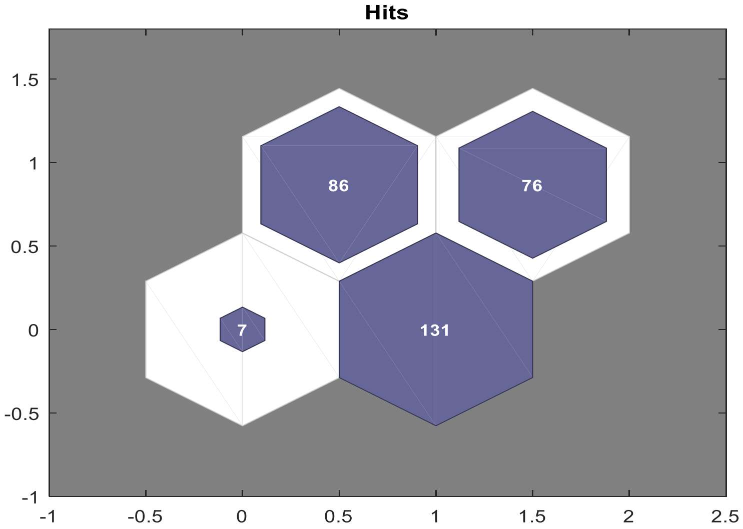

*Figure4: **SOM Sample Hits***

Figure4 shows data points associated with each neuron.

It is best if the data are fairly evenly distributed across the neurons. In our own case, the data are concentrated a little more in the lower-right neurons, but overall, the distribution is fairly even.

(b) **Credit Scoring Classifiers**

Table2: Credit Scoring Classifiers Performance Evaluation

MODELS	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
CART	96.3	97.1	95.7
SOM+CART	96.7	97.8	95.7

Table2 indicates that CART model has 96.3% accuracy, that is, the model gives 0.963 chance of correctly classifying the applicants in to their respective groups. Similarly, CART model has 97.1% Sensitivity, that is, the model gives 0.971 probability of correctly classifying the applicant as defaulter. Finally, CART has 95.7% specificity, that is, the model gives 0.957 probability of wrongly classifying applicant as a defaulter. SOM+CART model has 96.7% accuracy, that is, the model gives 0.967 chance of correctly classifying the applicants in to their respective groups. Similarly, SOM+CART model has 97.8% Sensitivity, that is, the model gives 0.978 probability of correctly classifying the applicant as defaulter. Finally, SOM+CART has 95.7% specificity, that is, the model gives 0.957 probability of wrongly classifying applicant as a defaulter.

(c) Visualization of CART and SOM+CART Models

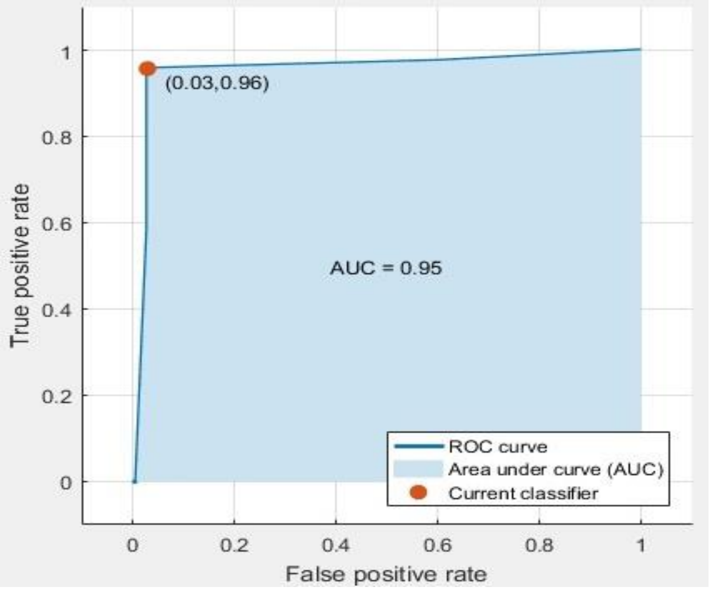

Figure5: CART ROC Curve

Figure5 shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR) of CART classifier. Since area covered under the curve is 0.95 and this is closer to the top-left corner of the curve indicating good performance by the model. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

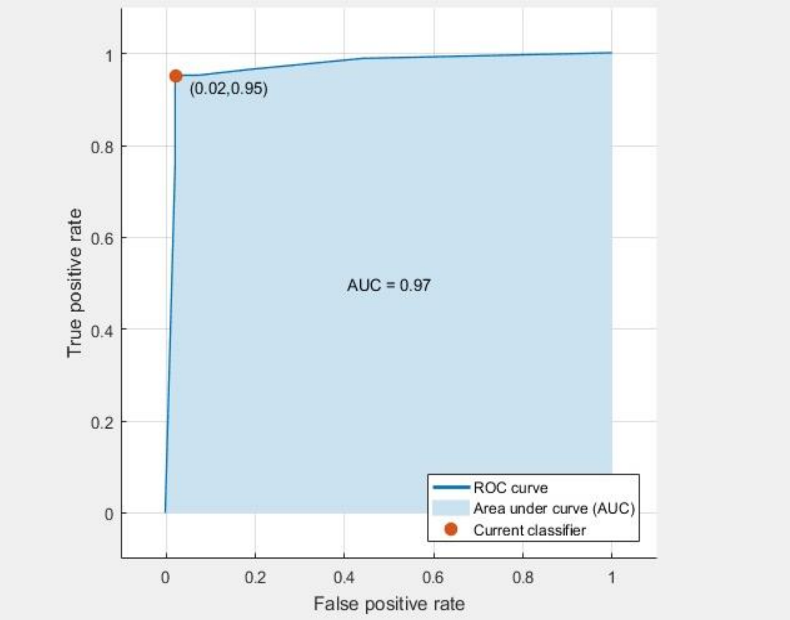

Figure6: SOM+CART ROC Curve

Figure6 above shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR) of SOM+CART classifier. Since area covered under the curve is 0.97 and this is closest to the top-left corner of the curve than that of CART ROC Curve indicating a better performance by the hybrid model.

Conclusion

Since credit loans and finances have risk of being defaulted, it is crucial for financial institutions to develop reliable credit scoring systems in order to predict creditworthiness of the borrowers. A standalone and two-stage techniques, SOM and SOM+CART were applied to predict whether the borrowers should be considered a good or bad credit risk. In this present work, a hybrid approach to building the credit scoring model using unsupervised learning based on self-organizing map (SOM) was proposed to improve the discriminant capability of classification and regression tree, to predict agricultural credit defaulters for Bank of Agriculture (BOA), Sokoto. Thus, the inputs of the two-stage models were the observations clustered in to four (4) groups by the SOM algorithm. The results indicate that the discrimination accuracy of the CART model can be improved from 96.3% to 96.7%. The results therefore indicate that the integration of SOM algorithm in to this technique made SOM+CART model to outperform CART. Since SOM algorithm improved the performances of CART in predicting Non-creditworthy applicants for BOA, Sokoto. It is recommended that future research work should be done considering other machine learning models hybridize with SOM algorithm.

References

Ali A., Ning C. and Bernardete R. (2016) "Improve credit scoring using transfer of learned knowledge from self-organizing map", *The Natural Computing Applications Forum* 2016.

Durand D. (1941) "Risk Elements in Consumer Instatements Financing". *National Bureau of Economic Research*, New York 1941. https://iq.opengenus.org/cart-algorithm/

Huysmans J., Baesens B. and Vanthienen J. (2005). "A comprehensible SOM-based Scoring System". *K.U. Leuven, Dept. of Applied Economic Sciences*, Naamsestraat 69, B-3000 Leuven, Belgium, School of Management, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

Kalamkas N. and Gulnar B. (2013). "Algorithmic Scoring Models". *Applied Mathematical Sciences*, Vol. 7, 2013, no. 12, 571 – 586.

Kohonen T (1982). "Self-organized formation of topologically correct feature maps". *Biological Cybernetics* 43:59–69.

Leo B., Jerome F., Charles J.S., and Olshen R.A. (1984). "Classification and Regression Trees". *Hall CRC* 1984.

Mohammad A. (2015). "Performance of Three Classification Techniques in Classifying Credit Applications into Good Loans and Bad Loans: A Comparison". *Department of Statistics, Uppsala University*.

Reza F. M. and Hermawan (2017). "Credit Scoring Using CART Algorithm and Binary Particle Swarm Optimization". *International Journal of Electrical and Computer Engineering (IJECE)*. Vol. 8, No. 6, December 2018, pp. 5425~5431.

Suleiman S. and M.S. Burodo and S. Issa (2017). "Credit Scoring using Principal Components Analysis-based Binary Logistic Regression". *Journal of Scientific and Engineering Research*, 2017, 4(12):99-110.

Suleiman S., Ibrahim A., Usman D., Yabo B.I. and Muhammad H.U. (2021). "Improving Credit Scoring Classification Performance using Self Organizing Map - Based Machine Learning Techniques". *European Journal of Advances in Engineering and Technology*, 2021, 8(10):28-35.

Vesanto J. and Esa A. (2000). "Clustering of the Self-Organizing Map". *Ieee Transactions On Neural Networks*, VOL. 11, NO. 3, MAY 2000.