Research Journal of Computer Systems and Engineering

ISSN: 2230-8571, 2230-8563

Volume 02 Issue 02 - 2021 (July to December)

Page 51:57

Ubiquitous Counter Propagation Network in Analysis of Diabetic Data Using Extended Self-Organizing Map

FRANKLIN ORE ARECHE
Bio-Engineering and Technology
UNIVERSIDAD NACIONAL DE HUANCAVELICA
https://orcid.org/0000-0002-7168-1742
franklin.ore@unh.edu.pe

Kartik Palsetty
Computer Science & Communication Engineering;Bio-Engineering and Technology;VLSI and Embedded Systems
Gh raisoni college of engineering
ORCHID ID-0000-0002-5533-6113
Kartik.Palsetty.ai@ghrce.raisoni.net

Article History

Received: 15 July 2021
Revised: 20 September 2021
Accepted: 22 November 2021

Abstract

Self-organizing maps are the most widely used methods to cluster and show data in scientific fields (SOMs). The better framework is counter-propagation (CPN), which has been successfully applied to many platforms including statistical analysis, pattern classification, and function approximation. The CPN method's collaboration with the Kohonen self-organizing map and classification network model makes it less error prone by a series convergence.In order to classify data from a diabetic database, this research proposed an enhanced SOM (E-SOM) with a decision tree that alters the ubiquitous counter propagation network (U-CPN) model. The aforementioned network, which uses a variety of learning rules, has a three layer network architecture.The input layer, the Kohonen layer, and the output layer make up the network's structure. However, the extended self-organizing map model trains both the Kohonen layer and the output layer using a modified Kohonen's learning algorithm.
Keywords: data processing, Self-organizing maps (SOMs), counter-propagation network (CPN), E-SOM, U-CPN, Kohonen's learning rule, classification precision.

1 Introduction:

Data mining is the extraction of abstracted data from the information and it is concept unknown of many data previously and use the data on the basis of round robin based information. Most of the data mining techniques are nearly close to the machine learning algorithm. Some of the data mining techniques are of statistical modeling and it sometimes represented as exploratory data analysis [1].The economic value is important for different values have been varied considerably. The identification of valuable is major challenge along the mode of transformation also the technique

implemented for data analysis [2].The remaining portions of the essay are structured as follows: The related work completed within the parameters of this study is covered in Section 2. The suggested work's technique is covered in Section 3. The findings and analysis of the model are covered in Section 4. Section 5 of the essay brings it to an end.

2 Literature review:

Work [3]studied on anomaly detection for online transactions using hybrid of Counter Propagation Neural Network and genetic algorithm (CPNN-GA). **Author [4]** designed a class of nonlinear dynamical methods utilizing Fuzzy Counter Propagation Neural Network (FCPN) controller. Work [5] discussed novel model in constructing a set of 541 compounds from Duluth database using counter propagation neural network. Work [6,7] explainedSOM and Multiple Back Propagation (MBP) in machine learning architecture along with its analysis for big biomedical data classification limitations. Authors in [8] determine self organizing map (SOM) predicting the threshold for optimal prediction model which helps in testing and labeling. Work in [9,10] describes a Batch-Learning Self Organizing Map with Weighted Connections avoiding false neighbor effects (BL-WCSOM).

3 Research methodology:

This section discuss about the proposed methodology. The input unsupervised classified data undergoes training with an extended model of self-organizing map (E-SOM) of neural networks and the supervised training data with less accuracy and then the data,the clustered data goes further with supervised classification algorithm Decision tree is the modified Counter propagation network model. Finally, pattern has been extracted and results are calculated. This section provides a detailed the description of projected architecture which is revealed in below figure 2.

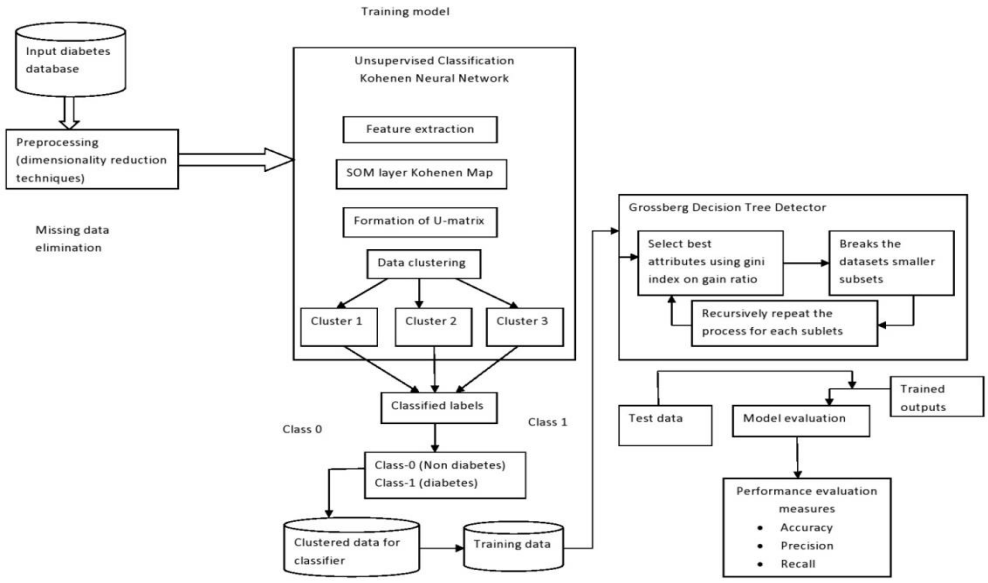

Figure 2: Proposed Architecture ESOM-UCPN

3.1 Pre-Processing By Dimensionality Reduction:

Statistical modeling let to the development of various Linear dimensionality reduction techniques in machine learning, and practically feasible to retrieve the primary data in pre-stage processing, and the above reduction techniques is most per-requisite tools for analyzing high dimensional noisy data. The process can be extensively used as optimization to solve the program to a complete a matrix manifold *M*, namely,

$$\text{Minimize } fX(M)$$

Subject to M ∈M

In a primary data the features has been extracted by the objective function $f_x(.)$is outlined,and the matrix manifold encodes by linear mapping P and it is determined by an equation $Y = PX^{-2}$ .All these techniques well-advised to specify as of two matrix and one of them as M .

3.2 Unsupervised classification with Extended SOM- Kohenen layer in U-CPN:

For this SOM algorithm, we consider n-input units for training vector and m-output neurons for number of clusters. As a topological map plane, the vector quantization technique has been represented in normal 2-D plot as a prototype by a dimensional view. Here each vector of input layer has been connected for trained weights to output neurons. For minimizing the distance between cells with their neighboring cells, the network topology distance calculation has been used in the proposed algorithm up to it reaches convergence and it update continuously as per weights of iteration. In maps unit the cell has been identified with the comparable samesets for primary data or clusters. Next in the training stage various clusters has been adjusted within the input data along with the variations in cell depending on their similarities.

3.3 PCA weights:

Recently, the most popularized dimensionality reduction techniques for pattern analysis is a PCA weights method it is a multidimensional-variable statistical method and it is furnished to determine the weight calculation. Its main task is calculating the weight of output nodes and limit the dataset dimension by taking the specific information and basically illuminating the dataset and scrutinizing the estimated value by eq. (1).

$Y = AX(1)$

$RY = AR_X A^T(2)$ (1)

X can be reconstructed as $X = A^T Y$. Instead of duplicating*X*using whole eigenvectors, it can be projected with fewer eigenvectors; thus, $\hat{X} = A_K^T Y$, where *k* eigenvectors are applied. In this case, the resulting weight of output nodes is given by eq. (2):

$$wgt = \sum_{i=1}^{d} \lambda_i - \sum_{i=1}^{k} \lambda_i = \sum_{i=k+1}^{d} \lambda_i \text{ [2]}$$

3.3.1 Initialize input nodes, output nodes, and connection weights:

N input vector is initialized to estimate U-matrix and the input is taken from the top (most frequently occurring) instances and establish a two-dimensional map (grid) of *M* output nodes. Initialize PCA weights values *wij*from*N* input nodes to *M* output nodes.

3.3.2 Present each set in order:

It depicts the instances as an input vector of N co-ordinates by eq. (3).

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2 \quad [3]$$

Where *xi*(*t*) assigns a value as 1 or 0 depends existence of *i*-th term in dataset conferred at time *t*.

3.4 Topology of U-CPN:

Ubiquitous Counter propagation network grouping two eminent techniques, the self-organizing map of Kohonen and the GDTD. The network acquires knowledge from the optimal look-up table and the mapping is the maximum optimization.Counter propagation network trains the data into twofold stages,ininitial stage the input vectors are grouped on to the Kohonen units.Clustering of the Kohonen units is mathematically presented by a dot product and it is the mathematical representation of two vectors and it does not depend upon the length of a vector.The above setbacks is overcomed by absolute value and it is defined as eq. (4)

$$D = \sum_{i=1}^{n} |x_i - v_i| \, [4]$$

In the above equation D quantify the position of x to weight vector v and it specify the distance between the input x to the vector v, the size of the Kohenen cluster layer is choosen dynamically and then the D value is choosenarbitarly depending of the application the value of D is choosen ,if the value of D is smaller then the networks is larger in size.

4 Performance analysis:

Below is an illustration of the performance analysis of the suggested method. Accuracy, precision, recall, and F1 score are the criteria that should be taken into account while evaluating a parameter. The topographical error rate and subsequent quantization error have been determined using the U matrix.

4.1 Dataset:

The Pima Indian Dataset (PID) was used in this study. The UCI Machine Learning Repository is where it was found. Basically, the National Institute of Diabetes, Digestive and Kidney Disease provided the original aggregate for this dataset. The PID dataset is built up with eight attributes, one output class, and a binary value to indicate whether or not a person has diabetes.There are 768 occurrences of the categorised traits, of which 500 are not diabetics and 268 are. PIMA has been chosen for this study because it is a prestigious and accepted standard dataset for comparing the effectiveness of methodologies across studies.

Table 3. Comparison of Performance of Proposed E-SOM system and Existing Algorithm(PID)

S.NO	**Performance Measure**	**Class 0**	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**	**Macro_AVg**
Techniques		**KNN**		**NN**		**ANN**		**Pro_ESOM**		
1	Accuracy	65	56	0.75	0.66	0.85	0.76	0.92	0.88	0.88
2	Precision	75	69	0.79	0.72	0.80	0.78	0.90	0.82	0.86
3	Recall	68	65	0.76	0.69	0.86	0.76	0.91	0.82	0.86
4	F1-score	65	58	0.76	0.64	0.81	0.72	0.91	0.82	0.86

Table 3 displays some of the observations from the Pima Indian Dataset (PID). From the instances in the diabetic datasets, the clustering result has been estimated. The instances with the same observation are then classified, and the performance measures of various KNN, NN, and ANN techniques are compared with the suggested Pro ESOM techniques.

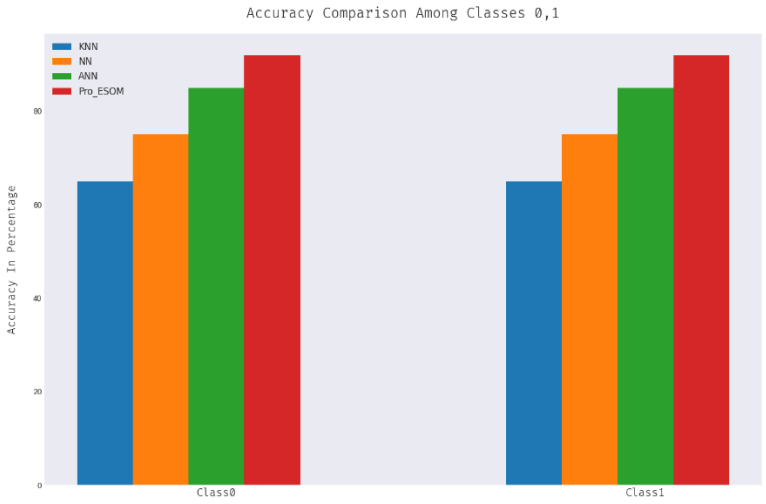

Figure 7: Accuracy comparison of classes 0,1 for ESOM (PID)

A graphic depiction of table 3 is shown in picture 7. The KNN technique, on the other hand, has produced the worst results, with a minimum Accuracy value of roughly 65% for Class 0 and 56% for Class 1. The NN model simultaneously gains greater accuracy compared to the prior one, which was roughly 75% for Class 0 and 66% for Class 1. As opposed to other existing methodologies, ANN steadily increases the Accuracy value to roughly 85% for Class 0 and 76% for Class 1. Finally, by achieving the maximum accuracy values of 92% for Class 0 and 88% for Class 1, the suggested Pro ESOM technique performs more effectively when compared to other models.

Figure 8: Precision comparison of classes 0,1 (PID)

The figure 8 is graphical representation for table 3. Whereas, KNN approach has resulted to worst performance by furnishing a minimum of Precision value of about 75% for Class 0 and 69 % for Class 1.Simultaneously, the NN model acquires more Precision compared to the previous one of about 79% for Class 0 and 72 % for Class 1. Whereas ,ANN gradually increase Precision value of about 80 % for Class 0 and 78 % for Class 1 than the other existing methods.Finally, the maximum precision values of 90% for Class 0 and 82% for Class 1 demonstrate that the suggested Pro ESOM approach performs more effectively than existing models.

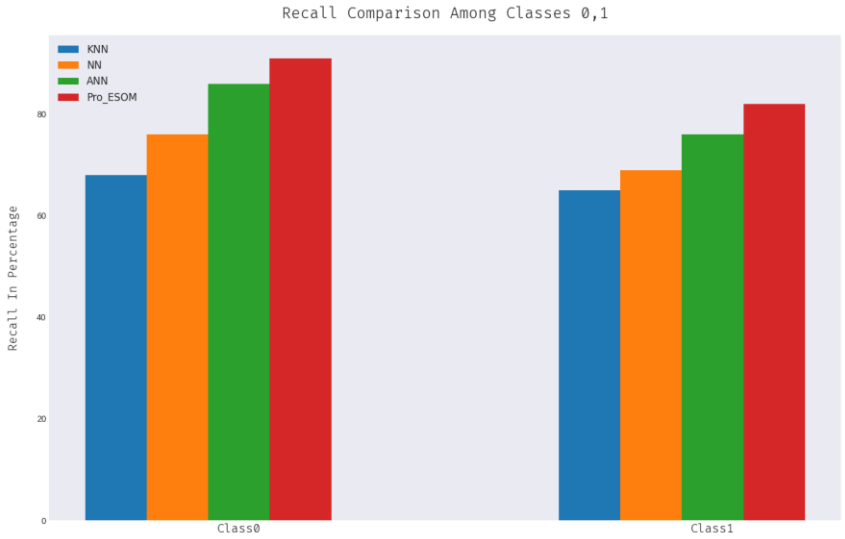

Figure 9: Recall comparison of classes 0,1 (PID)

The figure 9 is graphical representation for table 3. Whereas, KNN approach has resulted to worst performance by furnishing a minimum of Recall value of about 68% for Class 0 and 65 % for Class 1.Simultaneously, the NN model acquires more Recall compared to the previous one of about 76% for Class 0 and 69 % for Class 1. Whereas ,ANN gradually increase Recall value of about 86 % for Class 0 and 76 % for Class 1 than other existing techniques Finally, the suggested Pro ESOM approach achieves maximum Recall values of 91% for Class 0 and 82% for Class 1, demonstrating that it performs more effectively than other models.

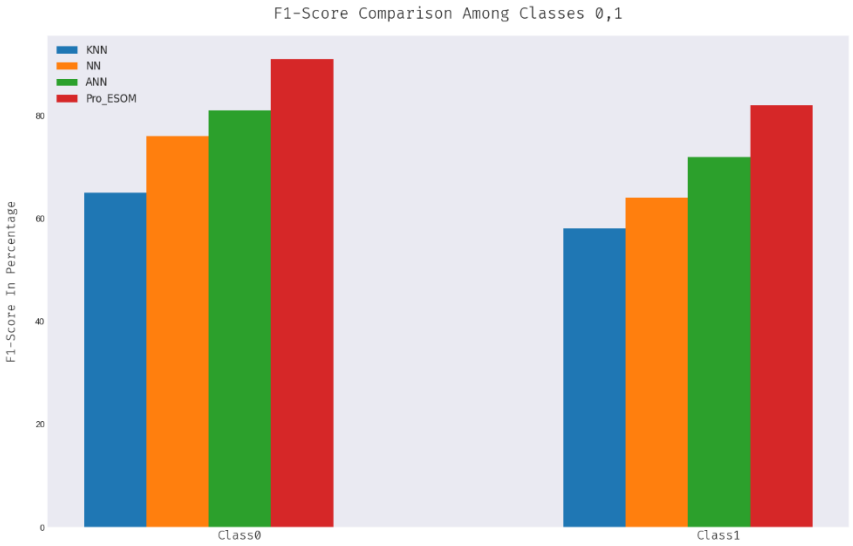

Figure 10: F1-Score comparison of classes 0,1 (PID)

The figure 10 is graphical representation for table 3. Whereas, KNN approach has resulted to worst performance by furnishing a minimum of F1-Score value of about 65% for Class 0 and 58 % for Class 1.Simultaneously, the NN model acquires moreF1-Scorecompared to the previous one of about 76% for Class 0 and 64 % for Class 1. Whereas ,ANN gradually increase the F1-Score value of about 81 % for Class 0 and 72 % for Class 1 than other existing techniques.Finally, the suggested Pro ESOM approach achieves the maximum F1-Score values of 91% for Class 0 and 82% for Class 1, demonstrating that it performs more effectively than other models.

5 Conclusion:

Research has integrated the two learning training algorithms ,namely, an extended SOM (E-SOM) model and a modified CPN (U-CPN) method. The diabetic data are applied to the unsupervised algorithm to segregates the data. Initially, the basic processing steps are applied to prepare to extract the feature from the attributes combinations and these process is used to initialize the two network architecture. The U-matrix of the E-SOM is used to analyze the various errors and then estimate the outcomes. Finally, the machine learning algorithm combines with the various initializing parameter to classify the diabetic data. This research analysis to depicts that U-CPN model has less error in classifying the diabetic data and has maximum learning speed but maximum in detecting the low positive rates compared to the negative rate during the training procedure than the E-SOM method.Overall U-CPN has higher classification accuracy for diabetic data.It might be argued that in order to get good results from supervised classification, the key parameters for the two architectures must be determined empirically and put to the test in experiments.

References:

[1] Saha, P. K., Patwary, N. S., & Ahmed, I. (2019, December). A widespread study of diabetes prediction using several machine learning techniques. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-5). IEEE.

[2] Abraham, A. (2018). Analysis of Genetic Algorithms and Distinctiveness in Back Propagation Neural Network on Diabetic Retinopathy classification.

[3] Wehrens, Ron, and Johannes Kruisselbrink. "Flexible Self-Organising Maps in kohonen 3.0." *Journal of Statis* (2018).

[4] Amusan, D. G., et al. "Hybrid Design using Counter Propagation Neural Network-Genetic Algorithm Model for the Anomaly Detection in Online Transaction." (2019).

[5] Sakhre, Vandana, et al. "Fuzzy counter propagation neural network control for a class of nonlinear dynamical systems." *Computational intelligence and neuroscience* 2015 (2015).

[6] Friedlander, David. "Pattern Analysis with Layered Self-Organizing Maps." *arXiv preprint arXiv:1803.08996* (2018).

[7] TahaniDaghistani, RiyadAlshammari "Diagnosis of Diabetes by Applying Data Mining Classification Techniques, Comparison of Three Data Mining Algorithms" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.

[8] Zeinali, Yasha, and Brett Story. "Structural impairment detection using deep counter propagation neural networks." *Procedia Engineering* 145 (2016): 868-875.

[9] Rawat, V. (2019). A classification system for diabetic patients with machine learning techniques. *International Journal of Mathematical, Engineering and Management Sciences*, *4*(3), 729.

[10] HUANG, S. X., YANG, Y. Y., LUO, Y. L., & CHEN, T. Y. (2018). Studies on cognitive model of type 2 diabetic nephropathy based on GA-BP neural network model. *Medical Journal of Chinese People's Liberation Army*, *43*(6), 483-489.