

## **Data Classification analysis using Distance Function in Learning Vector Quantization Algorithm**

Dushyant Purohit<sup>1</sup>, Yogendra Sharma<sup>2</sup>, Jitender Kumar<sup>3</sup>, Ritu Vashistha<sup>4</sup>

<sup>1</sup>Compucom Institute of technology & Management, Jaipur,

<sup>2</sup>Compucom Institute of technology & Management, Jaipur,

<sup>3</sup>Vivekananda Global University, Jaipur,

<sup>4</sup>Climax Softech Solutions, Jaipur,

**Abstract**— Artificial Neural Networks (ANNs) are human invented information processing objects, highly parallelized dynamic system that yields output by reapplying output as input. They have confirmed to be beneficial in solving problems which cannot be solved by using heavy algorithmic procedures in area like mining, categorization of large data, image processing etc. Proposed work focus on prediction of diseases like Cancer, Diabetes, patient hospitalization etc. Proposed work is to train the data with a new approach having combined supervised and unsupervised mode and get results with new conceptualize approach on test data. This approach is experimented by Learning Vector Quantization (LVQ) algorithm because of its characteristic of multimode working and uses heuristic approach for searching. Multimode working includes first is unsupervised step and another is supervised. Diabetes dataset is retrieved from UCI learning respiratory to test the proposed ideology on algorithm. The program is implemented on Matlab and result is shown using AUC (Area under the curve). Our approach is to improving accuracy in such a way that the performance may high and computational cost remain less.

**Keywords**—Clustering, Artificial Intelligence, Neural Networks, Self Organizing feature Map, Learning Vector Quantization, Voronoi vectors, Area under the curve(AUC)

### **I. INTRODUCTION**

Nowdays, in the automation world, major challenge faced by health care organizations is to provide quality services at reasonable prices. Quality service implies diagnosing patients properly and administering treatments that are effective. Quality less clinical diagnosis will cause fateful results unacceptable Clinical decisions area unit typically created supported doctors' suggestion and experience instead of on the data wealthy information hidden in the information. This observe results in unwanted biases, errors and excessive medical prices that affect the tandards of service provided to patients. This paper proposed integration of clinical support with computer-based patient records might scale back medical errors, enhance patient safety, decrease unwanted observe variation, and improve patient outcome [2]. Data classification needs selection of appropriate algorithm as every algorithm has own characteristic. Due to this main challenges, the computer program has to do filtration of data repetitively for perfect categoraziation. There are number of architectures in ANN designed for clustering. This work objective is to modify prediction of disease using ANN algorithm It is challenging task to reduce computational cost in heavy parallelized system The intention behind research project is reduce cost in ANN based system. This paper is organized as follows: Section 2 describes the literature studied for proposed diagnosis system using LVQ, Section 3 presents design of proposed methodology. Section 4 presents dataset used. Section 5 the simulation results and conclusions.

### **II. LITERATURE SURVEY**

As per theory, proposed in 1986 by Kohonen on Self Organizing feature Map, target of SOM is to encode a set of input vector  $\{i\}$  by finding a similar and smaller set of "prototypes."  $\{w_{1(x)}\}$  that results in good estimation to the input space provided to algorithm initially. This estimation is used by supervised learning. This is the basic idea of vector quantization which aims to process large input vectors to smaller set. Learning Vector Quantization algorithm (LVQ) is a enhanced version of SOM, where the trainer is available and used when we have tagged input data. The class information/estimation obtained from the prototype is used to give supervision so that LVQ can relocate Voronoi vectors. The motive behind is to make best classifiers decision reason. Verno vectors are basically, a space representation on  $x, y, z, w$  axis for classification and moving towards match class in space. For example, if we have a pattern that match for input vector let say  $x$ , and Verno vector let's say  $w$  satisfies, than vector  $w$  is moved towards the direction of input vector which is  $x$ , for current pattern. LQV is a two step process. The first step is feature choice/opting from input by applying SOM. Second step is to classify where to put this feature to individual classes [3,4,5]. The difference between one set to another set obtained by vector quantization is given by Euclidean measure is as

$$D = \sum_x \|X - (W_{1(x)})\|^2, \quad (1)$$

Where  $X$  = Input vectors

$W$  = Weight vector for output neuron (jth)

$D$  = Distance between vectors

According to previous papers, LVQ has been proposed with various special cases as below mentioned. Comparative study has been reported in [2] by presenting a intrusion detection model through hybrid system of LVQ improves the detection rate for dataset. Another approach presented in [3] is for optical Numeral Recognition with Principle Component Analysis (PCA). In this approach PCA perform better to produce coefficient vector of images. In the series of LVQ with special cases, recognition of speech is proposed in [4] by LVQ with Time-Delay Neural Network achieve high recognition rate. A comparative study between implementations of LVQ algorithm implemented using Euclidean distance function and Canberra distance function have been reported in [5], such as modified LVQ with Canberra distance function. Experimental results shows the improved method for distance calculation over traditional Euclidean Distance method *Fig. 2*, *Fig. 3* shows the simulation results proves the Canberra to perform better than conventional Euclidean function. In detail, Canberra distance function has been used in place of Euclidean distance method to find distance between points. Hybrid Evolution of genetic algorithm is also reported in [6] which use Particle Swarm Optimization for clustering. Proposed algorithm finds the refined initial centroid and provide geed optimization strategy with the drawback of more execution time as compared to k-means process.

### III. DISTANCE FUNCTION

This paper proposed analysis of data classification using metrices calculation in LVQ algorithm. Basically proposed ideology to enhance the prediction of diabetes using LVQ code. Algorithm has been implemented using the MATLAB 7.7.0 tool.

**LVQ METHODOLOGY :** In the data initialization block, algorithm load data from input files through commands and split the data. Data separation block include preprocessing of data Distance calculation is the CPU of an algorithm which are optimized to work within the condition. In this block, weight updation is performed and shortest distance between objects or data is found. This is done using traditional LVQ algorithm.

**CANBERRA DISTANCE FUNCTION :** This distance function is introduced and developed by G.N. Lance and W.T Williams in 1966 and 1967 respectively. When we need to find distance between pairs of point, where data is around the origin, in vector space, we use Canberra Distance function [5]. Canberra Distance function gives 2 outputs as TRUE only in case, if input to function is in rectangular form, and gives FLASE in case where the data is not in rectangular form or if some computational error occur during calculation,

$$D_{\text{Canberra}}(p, q) = \sum_{i=0}^n \frac{|p_i - q_i|}{|p_i| + |q_i|} \quad (2)$$

where,  $p$  and  $q$  are vectors for calculating distance.

**EUCLIDEAN DISTANCE FUNCTION :** This distance function is also known as Minkowski distance. When we need to find distance between points on straight line, we use Euclidean distance Euclidean Distance must bear the properties that all elements must lies on diagonal must be zero, and triangle inequality must be followed. This function does not reduce redundancies. The equation for Euclidean distance is as follow,

$$D_{\text{Euclidean}}(w, x) = d(x, w) = \sqrt{\sum_{i=1}^n (x_i - w_i)^2} \quad (3)$$

where,  $w$  and  $x$  are points in Euclidean space for calculating distance.

**CHEBYSHEV DISTANCE FUNCTION:** Chebyshev Distance is also known as maximum value distance and is computed as the absolute magnitude of the differences between coordinate of a pair of objects.

$$D_{\text{Chebyshev}}(p, q) = \max_i (|p_i - q_i|) \quad (4)$$

where,  $p$  and  $q$  are points or vector space for calculating distance.

### SIMULATION AND RESULT ANALYSIS

The experiment is carried out on Intel(R)core i3-4010U processor. The algorithm is tested on Pima Indian Diabetes, which having 768 instances of data and 9 attributes out of them one attribute listed in *Table I* is class attribute. 500 instances out of 768 instance belongs to class 0 and remain 268 belongs to class 1 [1] [5]. The training provided to system with 0.01 learning rate and 100 iterations. 90% of the data is used for training and rest of the data is used for testing.

| SNO | PARAMETERS(PIMA INDIAN DIABETES)                                                        |           |
|-----|-----------------------------------------------------------------------------------------|-----------|
|     | Recorded parameter                                                                      | Value     |
| 1   | Number of times patient pregnant                                                        | Numerical |
| 2   | Plasma-glucose concentration of Patient after 2 hours in an oral glucose tolerance test | Numerical |
| 3   | Patient's Diastolic blood pressure                                                      | Numerical |
| 4   | Triceps skin fold thickness (mm)                                                        | Numerical |
| 5   | 2-Hour serum insulin (mu U/ml)                                                          | Numerical |
| 6   | Body mass index of patient (weight in kg/(height in m) <sup>2</sup> )                   | Numerical |
| 7   | Diabetes pedigree function                                                              | Numerical |
| 8   | Age of patient                                                                          | Numerical |
| 9   | Class (positive or negative)                                                            | Numerical |

TABLE I.

**A. Receiver Operating Characteristic (ROC) comparison with diabetes dataset**

Fig.1. is the ROC curve plotted between TPR (Y-axis) and FPR (X-axis) for Canberra Distance function.

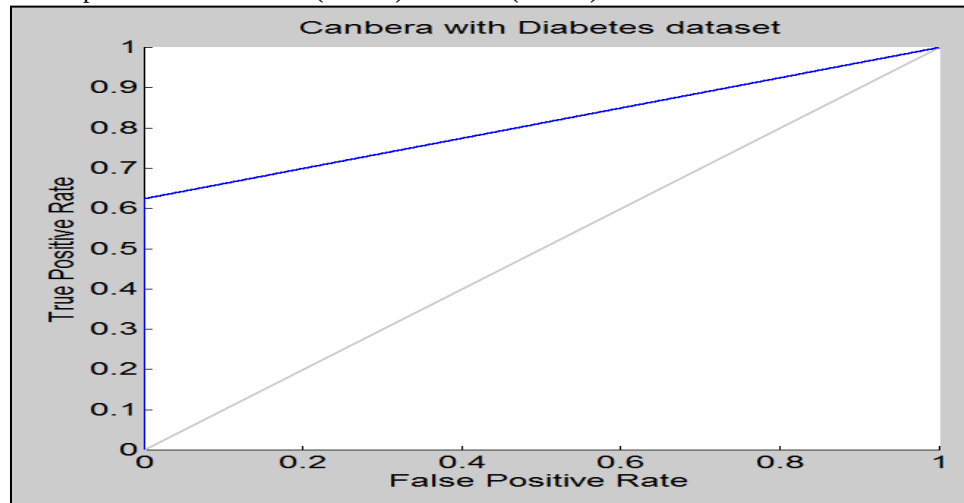


Fig. 1. Canberra Distance ROC curve on Pima India Dataset

Fig. 2. is the ROC curve plotted between TPR (Y-axis) and FPR (X-axis) for Chebyshev Distance function.

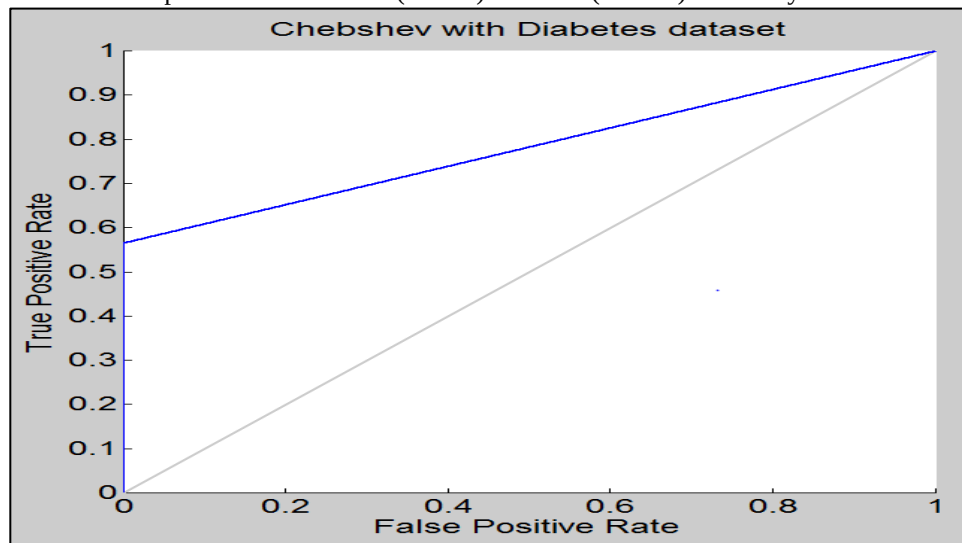


Fig. 2. Chebyshev Distance ROC curve on Pima India Dataset

ROC shows that chebyshev distance calculation method can also be a good option in place of traditional LVQ. In the below figures negative represents the negative cases, positive represents the positive cases identified by system, FNR in below figures despite the False Negative Rate and TPR despite the True Positive Rate.

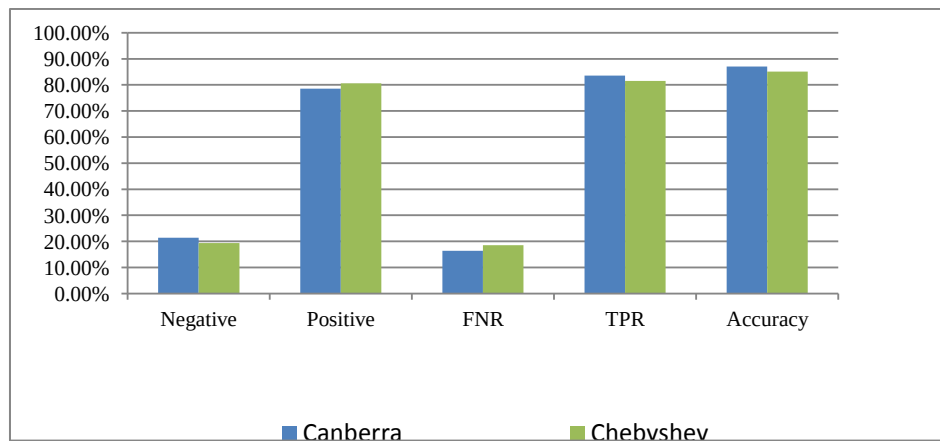


Fig. 3. Evaluation Result of confusion matrix on Pima Diabetes Dataset

#### IV. CONCLUSION

The result from the above simulation and bar graph proves that metrics evaluation over LVQ denoted with violet is more accurate. From above ROC, we conclude that change in methodology of core brain of traditional LVQ by Canberra is good as compared to Chebyshev distance function. **Hence we conclude that Chebyshev and Canberra is also good in place of Euclidean distance when classification is done on distributed data.** In the series of future improvement, sustainability analysis with respect to time of this research on big data, making of hybrid function for pattern classification, hybridization of different mining technique, speeding up will be done.

#### REFERENCES

- [1] Latha Parthiban, R. Subramanian, "Intelligent heart disease prediction system using canfis and genetic algorithm", in International Journal Of Biological, Biomedical & Medical Sciences, vol. 3, no. 3, 2008.
- [2] Reyadh Shaker Naoum, Zainab Namh Al-Sultani, "Hybrid system of learning vector quantization and enhanced resilient backpropagation artificial neural network for intrusion classification" in International Journal vol. 14 issue 2 pp 333-339, 2014.
- [3] Nishant Gupta, Arun Sharma, "A hybrid approach to optical numeral recognition (composition of pca & lvq)", in International Journal of Engineering and Research Technology, vol 2 Issue 1, January-2013.
- [4] S.A. Selouani, J. Caelen, "A hybrid learning vector quantization/timedelay neural network system for the recognition of Arabic speech", in Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), Antalya, Turkey, June 20-23, 1999.
- [5] Rehab F., Abdal-Kadar, "Genetically improved PSO algorithm for efficient data clustering", In Second International Conference On Machine Learning And Computing, pp. 71-75, 2010.
- [6] Y-T Kao, E. Zahara, and I-W. Kao, "A hybridized approach to data clustering," Expert Systems With Applications, vol. 34(3) pp. 1754-1762, 2008.
- [7] B.B. Firouzi, T. Niknam and M. Nayeripour, "A new evolutionary algorithm for cluster analysis," in Proc Of World Academy Of Sciences, Engineering And Technology, vol. 36 pp. 605-609, 2008.
- [8] T. Niknam, B.B. Firouzi, and M. Nayeripour, "An efficient hybrid evolutionary algorithm for cluster analysis," in World Applied Sciences Journal, vol. 4(2) pp. 300-307, 2008.
- [9] Li Xuan, Chen Zhigang, Yang Fan "Exploring of clustering algorithm on class imbalanced data" in The 8<sup>th</sup> International Conference On Computer Science And Education, pp. 89-93, 2013.
- [10] Sen Wu, Guiying Wei, "High dimensional data clustering algorithm based on sparse feature vector for categorical attributes," in IEEE, pp. 973-976, 2010.
- [11] Fahlman, S.E. "An empirical study of learning speed in back propagation networks". CMU-Cs, pp. 88-162, 1998.
- [12] F. Glenn, "A comprehensive overview of basic clustering algorithms" in Technical Report, University of Wisconsin, Madison. 2001.