Enhancement over Learning Vector Quantization through Distance Function

Preeti Jorwal[1], Vijeta Khicha[1], Vipin Jain[2]
[1]Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan Jaipur-302017, (INDIA)
[2]Department of Information Technology, Swami Keshvanand Institute of Technology, Management & Gramothan Jaipur-302017, (INDIA)
Email - priyanka1209sharma@gmail.com

Received 29.08.2019 received in revised form 18.12.2019, accepted 23.12.2019

Abstract : Artificial Neural Network (ANN) represents the scientific similarity between neuron of biological elements. These are computational models, which lightly stimulated by their biological equivalents. AI and ANNs are two stimulating and intertwined arenas in computer science. ANNs are hominid ready information handling systems that are grown up extensively in last thirty years. Researchers have used ANN for detection of different types of diseases. In this paper, we have proposed a hybrid function used with ANN for detection of cancer and diabetes diseases. Basically, this approach modifies the existing distance function and proposed a hybrid distance function. We have used this modified distance function for training and testing of model in supervised learning vector quantization for detection of diseases. The data sets have been taken form Medical Science for providing learning and examining. The various experiments were performed using MATLAB tool. The results show that the performance of enhanced ANN algorithm is far better than existing ANN for detection of cancer and diabetes diseases.

Keywords– Artificial Neural Network, artificial Intelligence, Learning Vector Quantization.

1. INTRODUCTION

For every specific problem, there is a different ANN configured. ANN is a type of neural network which challenges to intimate the way hominid brain works. Neural Network aims to get the human ability to adapt the change quickly and behave based on changes of data. The data may be result of researchers' efforts from market analysis and patient analysis. Neural network utilizes only raw data to predict the result of the problem like diagnosis report of patient, history of past sales and prices in market. ANNs is made up of interconnections between neurons called network architecture [1]. The training of an architecture aims to adjust the system and respond which is based on input provided to system. The architecture of network describes that how nodes are connected to each other and responsible for communication among nodes. Training is provided to the system for the setting of weight on them. The node with higher weight has more value and node with low weight is less valuable. These weights are responsible for finding the similarity and dissimilarity. The nodes acclimatize the pattern and find a match for it. The activation function is amenable for the output of neuron [2].

2. LEARNING VECTOR QUANTIZATION (LVQ)

LVQ is a supervised learning technique for clustering of statistical data, which uses the class information move slightly towards the 'Voronoi Vectors' so that, the quality of the classifier decision regions can be enhanced. LVQ aims to convert a large set of input vector {i} by detecting a smaller set of "prototypes."{w1(x)} which provides a good estimation to the original input space. This estimation is used by the supervised learning [3]. This is the elementary step of vector quantization aiming to process large input vectors to smaller set. This algorithm is a modified version of Self Organizing Map (SOM).

3. PROBLEM STATEMENT

The performance of Canberra distance function is limited to a dataset which is closer to zero. Hence, there are two limitations observed in Canberra distance function.

1. Use of Canberra distance is limited to datasets having the characteristic similar to features of Canberra distance measure.
2. It uses the only categorical data and there can be continuous data in some scenarios.
3. The simulation can be implemented by other data mining techniques in order to obtain better diagnostic outcomes.

4. METHODOLOGY

A discussion on a method to amend an ANN algorithm to improve their performance takes place in this chapter. It is well acquainted that choosing a best algorithm is usually a challenge [4]. The intention behind research work is to find the solution which shall be used in place of the distance function in pattern classification with LVQ. The Analysis work aims at conducting a comparative study between the implementations of Learning Vector Quantization algorithm implemented by using distance criteria, and replacement with a proposed hybrid method which is a proposed modification to the existing algorithm.

In this paper, work on LVQ algorithm by applying different distance functions and analyses of their accuracy has been focused. The second part of this paper is the application of Hybridization with the most widely used classification techniques used and that is the k-means algorithm. The first part of the proposed work uses medical dataset from UC Irvine Machine Learning Repository (UCI). The second part of the proposed work employs random data generated for pattern classification [5].

One of the most important issues in using the conventional Euclidean Distance is below described:

1. Euclidean may not perform well if data is at origin.
2. Euclidean is commonly used to evaluate the proximity of objects in two or three-dimensional space.
3. Canberra distance works only if data is at origin.

Thus, we propose Hybrid methods as possible substitutes for the conventional Euclidean distance function in pattern classification with LVQ.

The main objective of this methodology is to analyze the accuracy of Hybrid method with LVQ and k-means in Clustering problems, and evaluate them under specific conditions.

4.1 Distance Function and its Uses

The choice of distance function plays an important role in Categorization. The motive of any Categorization function is to contain integration of similar object from a group of homogeneous objects. The Categorization finds out objects which have the shortest distance from each other [6]. The choice of distance function is very difficult as there are many cases, where other distance function may give better outcomes of distance. So, it cannot be mentioned that distance function X is better than Y and distance function Z is more effective than There is another point to remember that while opting for distance function is there are continues and categorical both types of data. Distance measures are also depending on type of data. In Learning Vector Quantization, Euclidean distance function is used to measure the distance between the input vector and the weight vector.

4.2 Design of Hybrid Distance Function

This paper proposes a hybrid function for LVQ algorithm. The function is the integration of various distance functions, as its motive is to enhance the prediction of disease like Diabetes and Cancer using LVQ code, because LVQ can predict of diabetes with more accuracy, and also it is more sensitized. As per research, existing LVQ algorithm variants make use of Euclidean distance function and Canberra distance function in order to obtain the distance between the two vectors namely, the input vector and the weight vector.

4.2.1 Canberra Distance Function

This distance function is presented and developed by G.N. Lance and W.T Williams in 1966 and 1967 respectively. Canberra Distance function is used when the distance between pairs of point is needed to be found in vector space where data is nearby the origin.

Canberra Distance function gives 2 outputs as TRUE only in case, if input to function is in rectangular form, and gives FLASE in case when the data is not in rectangular form or if some computational error occurs during calculation [7].

$$\mathbf{D}_{\text{Canberra}}\,(\mathbf{p},\mathbf{q}) = \sum_{i=0}^{n} \frac{|p_i - q_i|}{|p_i|+|q_i|} \qquad \textbf{(i)}$$

Where, p and q are points in vectors space.

The Canberra distance not lies between 0 and 1. The term become unity if one of the coordinate is 0 and distance will not be affected. When both coordinates are 0, we need to be states as 0. The Canberra distance is extremely sensitive to a minor modification when both of the coordinate are close to 0.

4.2.2 Euclidean Distance Function

This distance function is also termed as Minkowski Distance. When we need to find distance between points on straight line, we use Euclidean distance Euclidean [8]. Distance must bear the properties and that all elements must lie on diagonal and it must be zero, and triangle inequality must be followed. The distance will be in concentric circles around the centre only when the sink is on the centre.

The Euclidean distance is calculated as follows:

$$\mathbf{D}_{\text{Euclidean}}\,(\mathbf{p},\mathbf{q}) = \mathbf{d}\,(\mathbf{q},\mathbf{p}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad \textbf{(ii)}$$

Where, p and q are points in Euclidean space.

4.2.3 Chebyshev Distance Function

Chebyshev Distance is also known as maximum value distance and is calculated as the absolute magnitude of the differences between coordinates of a pair of objects. The distance will be in concentric squares around the centre only when the sink is on the centre.

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

Where, p and q are points or vector space.

Suppose we have a coordinate value of two objects denoted with O say A and B and features denoted with F1 to F5 as shown in table-1 below the object that doesn't possess features is dictated as (Not Available) NA.

Table 1: Objects with features

F/O	F1	F2	F3	F4	F5
A	0	3	4	5	NA
B	7	6	3	1	NA

Both objects A and B have 4 features.

Now, the Euclidean distance (Dab) between object A and object B is

$$D_{ab} = \sqrt{(0-7)2 + (3-6)2 + (4-3)2 + (5+1)2}$$

$$= \sqrt{49+9+1+36} = 9.747$$

Chebyshev Distance (D_{ab}) between object A and object B is

$D_{ab} = \max\{|0\text{-}7|, |3\text{-}6|, |4\text{-}3|, |5\text{+}1|\} = \max\{7, 3, 1, 6\} = 7$

Canberra Distance (D_{ab}) between object A and object B is

$$D_{ab} = \frac{0-7}{0+7} + \frac{3-6}{3+6} + \frac{4-3}{4+3} + \frac{5+1}{5+1} = 2.467$$

4.3 Proposed Hybrid Model

In our proposed distance function, we have found that the farthest distance from Canberra, Chebyshev, Euclidean. This function is applied after computing distance from

$D_{max1} = \max(d_{Canberra}, d_{Chebyshev})$

$D_{max2} = \max(d_{max1}, d_{Euclidean})$

$D_{Hybrid} = \max(d_{max2})$

The farthest distance from point p to point q in vector space through Hybrid Distance formula is given by above equation. The hybrid distance in above example will be

$D_{max1} = \max(9.747, 7) = 7$

$D_{max2} = \max(7, 2.467) = 2.467$

$D_{min2} = \max(7, 2.647) = 2.46$

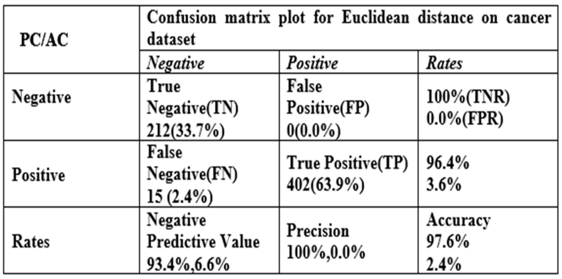
Algorithm: -Input: Data set, point p and q.

Output: Maximum distance between two points.

1. Calculate distance by Canberra Distance function, Euclidean distance Function and Chebyshev Distance Function.
2. Obtain the minimum from equations between resultant distances from step 1.
3. Again obtain the minimum from above equation. The resultant distance will be minimum distance between point p and q.

Here, we simulate and analyze the performance of algorithm with conventional Euclidean, Canberra, Chebyshev and Hybrid method as proposed method. A Confusion matrix and ROC curve have been plotted for algorithm with each distance function.

Table 2: Confusion Matrix for Euclidean distance with Cancer Dataset



PC/AC	Confusion matrix plot for Euclidean distance on cancer dataset		
	Negative	*Positive*	*Rates*
Negative	True Negative(TN) 212(33.7%)	False Positive(FP) 0(0.0%)	100%(TNR) 0.0%(FPR)
Positive	False Negative(FN) 15 (2.4%)	True Positive(TP) 402(63.9%)	96.4% 3.6%
Rates	Negative Predictive Value 93.4%,6.6%	Precision 100%,0.0%	Accuracy 97.6% 2.4%

As per Confusion Matrix plot retrieved for Euclidean Distance function, below mentioned parameter values are observed:

1. Negative Cases: 212 (212+0) out of 629 (33.7%)
   TNR = TN / TN + FP = 212/212which is100%.
   FPR = FP / TN + FP= 0/212 = 0.0%.
2. Positive Cases: 614(212+402) cases of 629 (97. 3%)
   FNR = FN / FN + TP = 15/417 = 3.59%
   TPR = TP / TP + FN = 402/417= 96.40 ~ 96 %
4. Precision (P) =TP / FP + TP
   = 402/ (0+402) = 100%
   Accuracy (AC) = TN+TP/TN+FP+FN+TP
   = (212+402) / (212+0+15+402) = 614/629= 97.6% ~ 97.6%.

Here TPR rise up to 0.96.

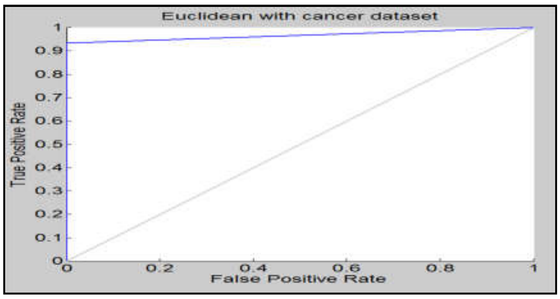


Figure 1: Euclidean Distance ROC curve on cancer Dataset

Table 3: Confusion Matrix for Canberra distance with Cancer Dataset

PC/AC	Generated confusion matrix with cancer dataset for Canberra distance		
	Negative	*Positive*	*Rates*
Negative	True Negative(TN) 225(35.8%)	False Positive (FP) 0(0.0%)	100%(TNR) 0.0%(FPR)
Positive	False Negative(FN) 2 (0.3%)	True Positive(TP) 402(63.9%)	99.5% 0.5%
Rates	Negative Predictive Value 99.1%,0.9%	Precision 100%,0.0%	Accuracy 99.68% 0.32%

As per Confusion Matrix plot retrieved for Canberra Distance function, below mentioned parameter values are observed:

1. Negative Cases: 225 (225+0) out of 629 (35.8%)
   NR = TN / TN + FP = 225/225 which is 100%.
   FPR = FP / TN + FP = 0/225 = 0.0%
2. Positive Cases: 627(225+402) cases of 629 (99.68%)
   FNR = FN / FN + TP =2/404 = 0.4%
   TPR = TP / FN + TP = 402/404= 99.50 ~ 99.50 %
3. Precision (P) = TP/TP+FP = 402/ (402+0) = 100%
4. Accuracy (AC) = TN+TP/TN+FP+FN+TP
   = (225+402) / (225+0+2+402) = 627/629
   = 99.68% ~ 99.7%

Here TPR (0.99) rise up to more than TPR of Euclidean Distance ROC which is 0.9 at x-axis. It shows Canberra Distance Function works more accurately compared to Euclidean Distance.

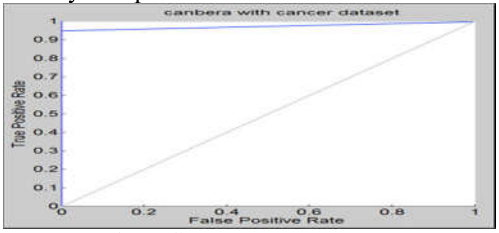


Figure 2: Canberra distance ROC curve on cancer Dataset

Table 4: Confusion Matrix for Hybrid distance with Cancer Dataset

PC/AC	Generated confusion matrix with cancer dataset for Hybrid distance		
	Negative	*Positive*	*Rates*
Negative	True Negative (TN) (226)35.9%	False Positive (FP) 0(0.0%)	100 % (TNR) 0.0%(FPR)
Positive	False Negative (FN) 1 (0.2%)	True Positive (TP) (402) 64.1 %	99.75%(TPR) 0.25 (FNR)
Rates	Negative Predictive Value 99.3%, 0.7%	Precision 100%,0.0%	Accuracy 99.84%,0.16 %

As per Confusion Matrix plot retrieved for Hybrid Distance function, below mentioned parameter values are observed:

First Observation for the function resultant confusion matrix gives result equivalent to Canberra Distance Function. From second observation onwards obtained confusion matrix as shown above. Calculation is descripted based on above matrix below

1. Negative Cases: 226 out of 629 = 35.9
   TNR = TN/TN+FP = 100%.
   FPR = FP/FP+TN = 0.0%
2. Positive Cases: 628(402+226) out of 629 cases (99.84%)
   FNR = FN/FN+TP
   =1/403 = 0. 2%
   TPR = TP / FN + TP
   = 402/1+402 = 99.75%
3. Precision (P) = TP/TP+FP
   = 100%
4. Accuracy (AC) = TN+FP/TN+FP+FN+TP
   = (628) /(629)=99.84%

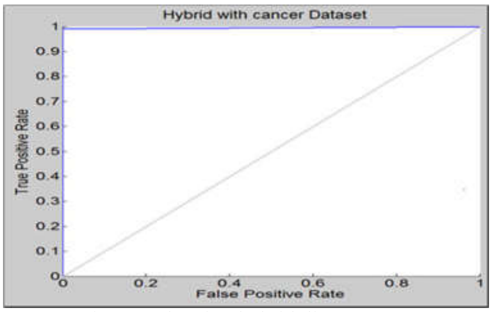


Figure 3: Diagram showing hybrid distance ROC curve on cancer Dataset

Table 5: Confusion Matrix for Chebyshev Distance with Cancer Dataset

PC/AC	Confusion matrix plot for Chebyshev distance on cancer dataset		
	Negative	*Positive*	*Rates*
Negative	True Negative (TN) 203(32.5%)	False Positive (FP) 0(0.0%)	100%(TNR) 0.0%(FPR)
Positive	False Negative (FN) 24 (3.8%)	True Positive (TP) 402(63.9%)	94.6% 5.6%
Rates	Negative Predictive Value 89.4.0%,10.6%	Precision 100%,0.0%	Accuracy 96.2% 3.8%

As per Confusion Matrix plot retrieved for Chebyshev Distance function, below mentioned parameter values are observed:

1. Negative Cases: 203 (203+0) out of 629 (32.27%)
   True Positive Rate(TPR) = TN/TN+FP = 203/203 which is 100%.
   False Positive Rate(FPR) = FP/FP+TN = 0/203 = 0.0 %.
2. Positive Cases: 605(203+402) cases of 629 (96.18%)
   FNR =FN/FN+TP = 24/426 = 5.63%
   TPR = TP/ FN+TP = 402/426= 94.36 ~ 94 %
3. Precision (P) = TP/ FP+TP = 402/ (0+402) = 100%
4. Accuracy (AC) = TN+TP/TN+FP+FN+TP
   = (203+402) / (203+0+24+402)
   = 605/629= 96.18% ~ 96%

Here TPR (0.94) is less than TPR obtained from conventional Canberra Distance Function and Euclidean Distance Function.

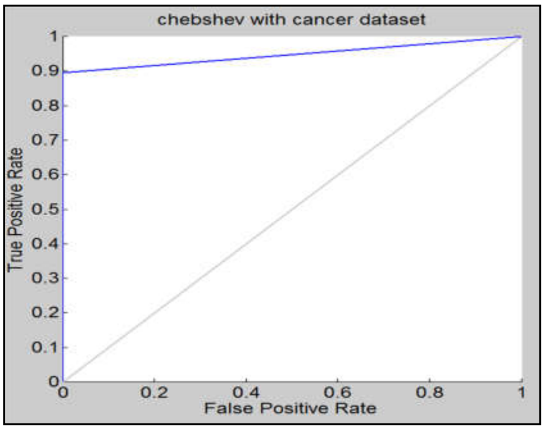


Figure 4: Chebyshev distance ROC curve between TPR and FPR on cancer Dataset

4.4 Hybridization over k-means algorithm

Here, we apply hybrid on various other different metrics and calculate centroid between each applied distance function and hybrid distance function, which replaces the total squared distance function. The outcome distance, which has been shown, is the distance retrieved from the hybrid function and it is the shortest one as compare to Euclidean, Canberra, Chebyshev. The proposed work gives better results and a better way than existing methods of single function result.

Hybrid Version

K-Means algorithm is used for clustering /quantization of input data. We cluster the outcome of the K-Means algorithm using the hybrid distance function in K-Mean algorithm, and distance between the two centroids from the equivalent clusters is also verified to gain the shortest distance.

Algorithm:

Notations: - D-distance calculated by matrices.

Input: V1, V2 – Point between these two the distance is calculated.

1. Calculate distance with each distance metrics among V1 and V2.
2. Calculate minimum of d among all distance matrices.

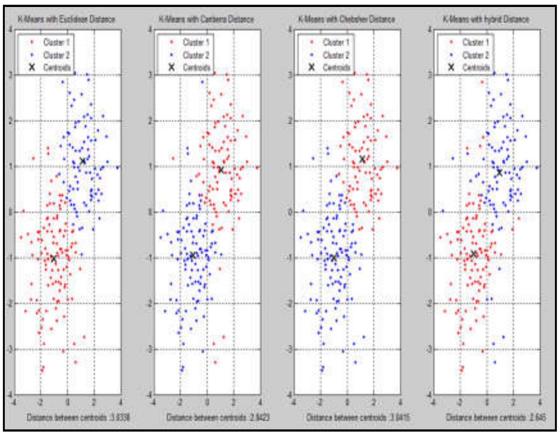


Figure 5: Cluster distribution using K-Means

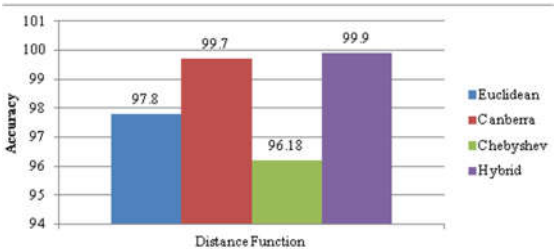


Figure 5: Outcome of the distance functions by bar graph

5. CONCLUSION

The contribution of this work is that a new solution of all distance function can be made apart from their application area (dimension, area) which gives more precise result, because Canberra distance function proves best in cases where values taken for a data set are nearer to zero. Therefore, its procedure is limited to the data sets only having categorical data. This is limitation. The proposed Hybrid LVQ performance is enhanced up to 90.6% from 87.1% of Canberra Distance.The continuous data requires in some diagnostic cases. Here, technique was used in the proposed work to test the sustainability of algorithm proposed with hybridization of distance function

REFERENCES

[1] Anderson, J (1995), An Introduction to Neural Networks: MIT Press

[2] Muhammad A. Sapon K.I (2011), " Prediction of Diabetes by Using Artificial Neural Network", International Conference on Circuits system and simulation.

[3] Paul S. Heckerling, Gay J. Canaris, Stephen , Flach, Thomas G. Tape,Robert S. Wigton, Ben S. Gerber, ―Predictors of urinary tract infection based on artificial neural networks and genetic algorithms,‖ international journal of medical informatics 7 6, 2007.

[4] Deza E., Michel M (2009). Encyclopedia of Distances. Springer. p. 94.

[5] Heon Gyu Lee, Ki Yong Noh, and Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis Using Linear and Nonlinear Features of HRV", T. Washio et al. (Eds.): PAKDD 2007 Workshops, LNAI 4819, pp. 218–228, 2007, © Springer-Verlag Berlin Heidelberg 2007.

[6] Bayu Adhi Tama, Rodiyatul F.S., Hermansyah, "An early detection method of Type-2 Diabetes Mellitus in Public Hospital", TELEKOMNIKA, ISSN 1693-6930 Vol.9 No.2 (2011), pp. 287-294.

[7] Hand, David J.(2009); Measuring classifier performance: A coherent alternative to the area under the ROC curve, Machine Learning, Vol 77 pp 103–123

[8] Gonzalez A.I., Grana M. Anjou A.D., "An analysis of the GLVQ algorithm", Department de Economia de la Excma. Diputación de Guipuzcoa, and project PGV9220 of the Gobierno Vasco

[9] E. Mwebaze, P. Schneider et al., "Divergence based Learning Vector Quantization", ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

[10] Biehl M, Bunte K, Schneider P (2013), "Analysis of Flow Cytometry Data by Matrix Relevance Learning Vector Quantization". PLoS ONE 8(3): e59401. doi:10.1371/journal.pone.0059401

[11] Blanca S.Leona, AlmaY. Alanisb,n, EdgarN. Sancheza, Fernando Ornelas-Tellezc, EduardoRuiz-Velazquezb, ―Inverse optimal neural control of blood glucose level for type1diabetes mellitus patients, Journal of the Franklin Institute 349 (2012) 1851–1870.

[12] Ravi Sanakal, Smt. T Jayakumari, ―Prognosis of Diabetes Using Data mining Approach- Fuzzy C Means Clustering and Support Vector Machine,‖ International Journal of Computer Trends and Technology (IJCTT) – volume 11 No 2 May 2014.

[13] Veena Vijayan V. Aswathy Ravikumar, ―Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus, International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014

[14] M. Durairaj, V. Ranjani, ―Data Mining Applications In Healthcare Sector: A Study, international journal of scientific & technology research volume 2, issue 10, October 2013.