icrrd
ICRRD JOURNAL

INTERNATIONAL CENTER FOR RESEARCH AND RESOURCES DEVELOPMENT article

ISSN Number: 2773-5958, SSM Number: (1347468-T), doi.org/10.53272/icrrd, www.icrrd.com

Application of Information Gain Based Weighted LVQ for Heart Disease Diagnosis

Radhanath Patra

Research scholar, Electronic Science, Berhampur University, Odisha 760007, India

**Corresponding author; Email: 1radhanath.patra@gmail.com*

Received: 12 June 2022	**Available Online**: XX XX 2022
Revision: 29 July 2022	**Published:** XX XX 2022
Accepted: 16 August 2022	**Volume-3, Issue-3**
✓ **Cite This:** *ICRRD JOURNAL,* 2022, 3(3), 1XX-1XX	

ABSTRACT: In this paper, an Information Gain based Weighted Linear Vector Quantization (IG-WLVQ) is applied to heart dataset available in UCI machine learning repository for prediction of heart disease. It considers all attributes of the data set. The IG-WLVQ method weights the attributes according to their information gain while training the dataset. It is found that the classification accuracy approaches to 98.9%.

Keywords: LVQ, IG-WLVQ, ML

1. Introduction

The mortality rate due to heart disease is increasing day by day in human beings. It is a matter of serious concern worldwide. Therefore, effective measures are highly essential to control the disease. Machine learning techniques have already proven as the most reliable and perfect platform for health care sector. The Cleveland heart dataset of UCI machine repository is considered for this analysis. Most of the researchers considered 13 attributes and one class level of Cleveland heart dataset neglecting the significance of remaining attributes [1][2]. The Majority of research which was carried out with those attributes provides good accuracy due to small data set. Some of the authors use ensemble technique for feature extraction to improve classification accuracy [3][4]. In this paper, an attempt has been made to use the whole dataset without any information loss. **Our unique purpose is to use all attributes which is applied to LVQ [5].** This must be accomplished by placing information gain along with LVQ algorithm such that the classifier performance is boosted recognizing it as a good classifier.

Related work:

Abdullah Caliskan and Mehmet Emin Yuksel(2017) implemented deep neural network classifier model with two encoder and a softmax layer to analyze the coronary artery disease medical dataset by taking 303 records and 14 attributes including class level. The classification accuracy was 87.64%[6]. Ashraf et al. (2019) also used deep learning neural network for the same dataset and the classification accuracy was found to be 95%[7]. KaanUyar and Ahmet ilhan(2017) developed genetic algorithm based recurrent fuzzy neural network(RFNN) and achieved classification accuracy of 96.63%[8]. Poornima v and Gladis D (2018) proposed orthogonal local preserving projection (OLPP) and hybrid classifier technique. In OLPP techniquethe authors implemented principal component analysis followed by orthogonal basis vector for dimension reduction of original dataset and afterwards a hybrid classifier composed of group search optimization algorithm along with Levenbergmarqartdt algorithm to get the best classification accuracy of 94% [9]. AmitaMalav and KalyaniKadam (2018) used knn clustering with multiplayer perceptron technique and achieved classification accuracy of 93.52% [10]. Mutasem Sh.Alkhasawneh (2019) proposed a hybrid neural structure, a combination of forward neural network with ELMAN neural network (HECFNN) for classification of the same dataset along with other four datasets of UCI machine learning repository. The neural network structure consists of input layer, hidden layer, context layer and an output layer. The classification accuracy achieved with the structure is 94.01% [11]. Ali et al. (2019) introduced chi^2 ($\lambda^{2)}$ test to remove noisy features and applied deep neural network (DNN) to the extracted dataset of 303 records with 76 attributes. The proposed system was designed to solve the over fitting and under fitting problem occurred due to various machine learning approach. The classification accuracy was found to be 93.3% [12].

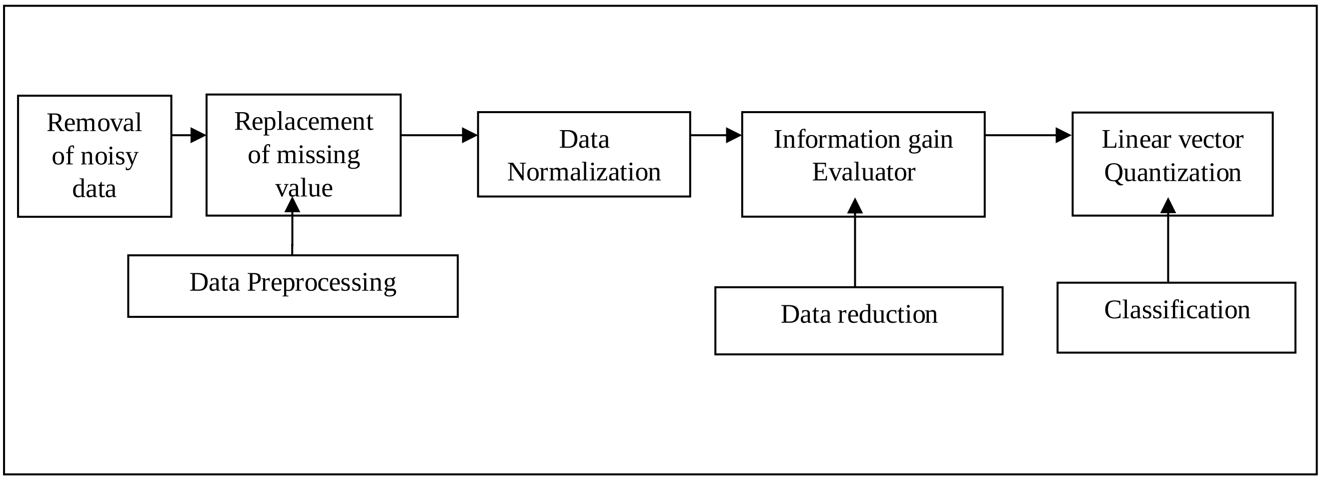
Kathleen H Miao and Julia H Miao (2018) developed an enhanced deep learning model with the same dataset by taking 28 attributes which resulted in an accuracy of 83.6%.[13].

2. Proposed Methodology/Algorithm:

The UCI machine learning Cleveland heart dataset consisting of all 303 records having 76 attributes, is considered in the proposed system.

The basic function of the classification problem is presented in fig. 1. It involves three operations as follows:

i. Data preprocessing
ii. Feature Selection/Information Gain Evaluator
iii. Classification



f proposed IG-WLVQ classification

i) Data Preprocessing:

The first phase is data preprocessing where redundant and insignificant attributes such as columns having constant are discarded [14].

To preserve essential attributes, missing data of a column are replaced with their corresponding mean value [15] [16]. Then normalization process begins. This represents all continuous values in form of binary form within range of 0 to 1. With this approach, we have 51 attributes, and among that one binary class label representing presence or absence of heart disease.

$$Normalization = \frac{(x_i - meanvalueofcoloumn)}{(maximumvalueofcolumn - minimumvalueofcoloumn)}$$

$$i = 1,2,3 \ldots \ldots n$$

ii) Feature selection/Information gain evaluator:

In the proposed method, information gain of filtering technique is used to find the weight of attributes. Information gain is one of the feature extraction principle in which attributes having more information are evaluated [17]. Information gain helps to determine the content of information present in the attribute. The attributes with non-zero information gain are considered for processing. The information gain is calculated as follows:

Let $x_i = \{x_1, x_2, x_3 \ldots \ldots x_n\}$

Entropy $H(x) = E[I(X)]$

$$H(x) = -\sum_i^n p(x_i) log_2 p(x_i)$$

Conditional Entropy

IF X and Y are represented two values of x_i and y_j

Then $H[X/Y]$=$\sum_{i,j} p(x_i, y_j)\, log\left(\frac{p(x_{i,}y_j)}{p(j)}\right)$

$$IG_i = \left(H(x) - H(X/Y)\right)$$

iii) Classification:

Linear vector quantization technique is based on the principle of winner take all algorithm concept. LVQ algorithm is a simple machine learning approach but its application and flexibility makes it one of the powerful classification techniques under supervised learning. The power of LVQ lies on the Euclidean distance [18]. In our research paper, a weighted LVQ is used for classification of heart disease. The distance is calculated based on weightage of the information of the corresponding attributes. This helps in preserving those attributes having less information gain. The pseudo code for the IGW-LVQ is as presented below:

Pseudo code of IGWLVQ

X_i =Training vector where i=1 to n, $\{X_1,X_2,X_3,X_4, \ldots X_n\}$

T=Class for training vector X_i

w_j=Weight vector for j^{th} output unit

c_j=class associated with j^{th} output unit

Step1: Find the number of class label and consider as m.

Step2: Choose the weight vector corresponding to m unit and assign the class label.

Such that w_j where j= 1to m.

Step3: Initialize the alpha value=α

Step 4: Find the information content of each attributes

Step5: Continue with step 6 to 9 till the stopping condition is not achieved.

Step5: Calculate the Euclidean distance for j=1 to m and I =1 to n.

D (i, j) =$\sqrt{(x_i - w_j)^2 * IG_i}$.

Step6: Obtain the winning unit where d (j) has minimum value

Step7: Calculate the new weight of the winning unit by the following relation

(i) If T= c_j Then $w_j = w_j(old) + \alpha[x - w_j(old)]$

(ii) if T≠ c_j Then $w_j = w_j(old) - \alpha[x - w_j(old)]$

Step 8: Reduce the learning rate α

Step9: Test stopping condition.The stopping condition is either maximum number of iterations or sufficiently small value of the learning rateα.

Figure 2. Shows the flow-chart for proposed algorithm.

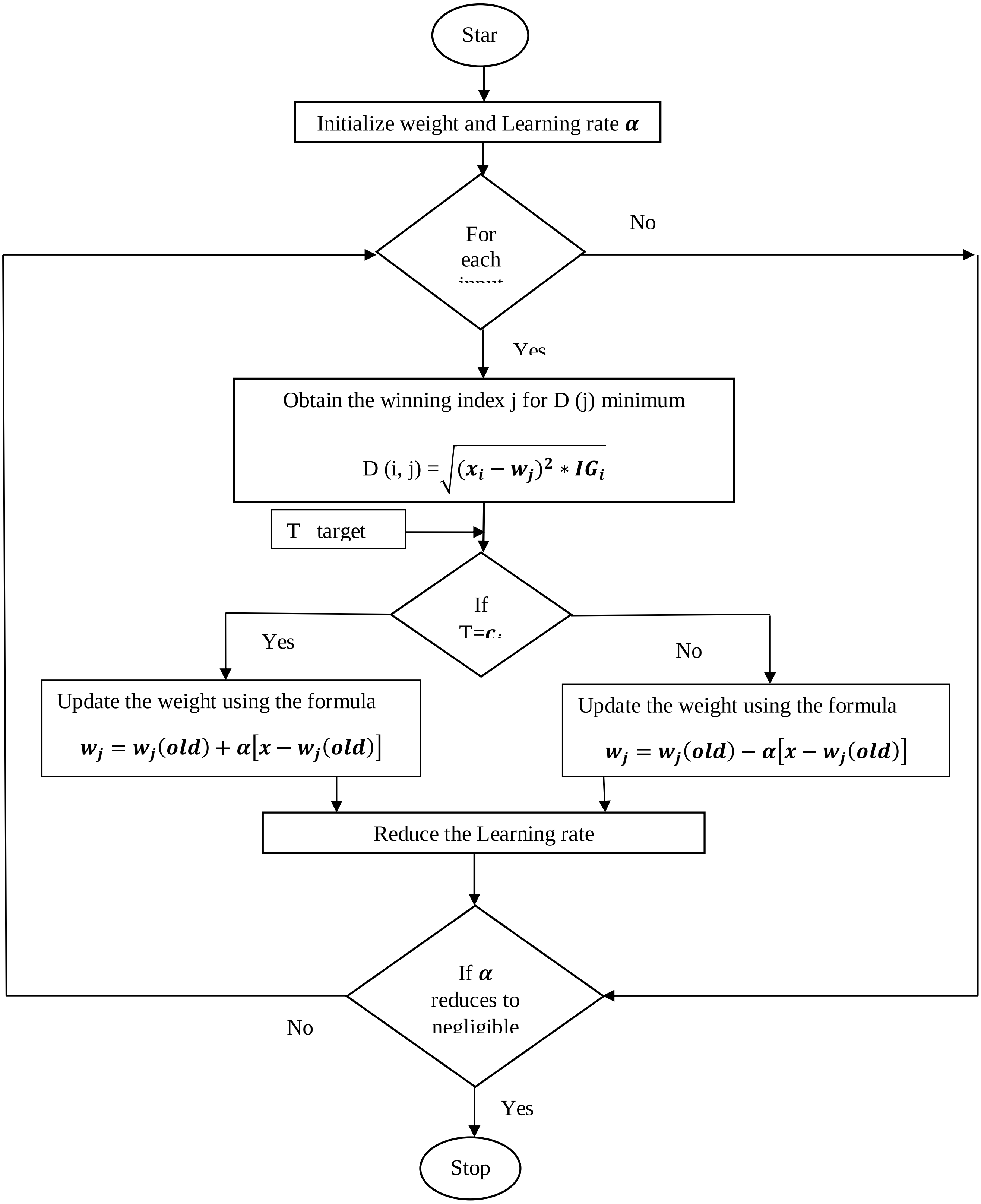

Star
Initialize weight and Learning rate α
For each
No
Yes
Obtain the winning index j for D (j) minimum
D (i, j) = $\sqrt{(x_i - w_j)^2 * IG_i}$
T target
If
Yes
No
Update the weight using the formula
$w_j = w_j(old) + \alpha[x - w_j(old)]$
Update the weight using the formula
$w_j = w_j(old) - \alpha[x - w_j(old)]$
Reduce the Learning rate
If α reduces to negligible
No
Yes
Stop

Fig2: Flowchart of IG-WLVQ Classification

Result Analysis:

After preprocessing the dataset has 303 records with 50 attributes and a class label. In the proposed technique, euclidean distance is multiplied with information gain in linear vector quantization. Thus it developed a new method of feature selection process based on weightage of information gain. The attributes depending up on the rank of information gain played a key role to change value of weight which we have considered in our LVQ process. These weight values are selected randomly and as iteration proceeds the selected weight values are updated and optimized which is used for testing. The learning parameter alpha is set to 0.2. It results 98.9% classification accuracy. On the other hand, the classification accuracy by standard LVQ for the same dataset is found to be 90%.

Sl.No.	Standard LVQ	IGLVQ(information gain with LVQ)	Improvement
1	70	76.66	6.66
2	73.33	83.33	10.00
3	76.66	80	3.34
4	80	86.66	6.66
5	83.33	86.66	3.33
6	86.66	90	3.34
7	90	98.9	8.9

Table 1: Accuracy measurement at various iteration in-terms of percentage

Average increase in accuracy of 5.6% is achieved in implementation of IGLVQ over standard LVQ by simulating the two models seven times.

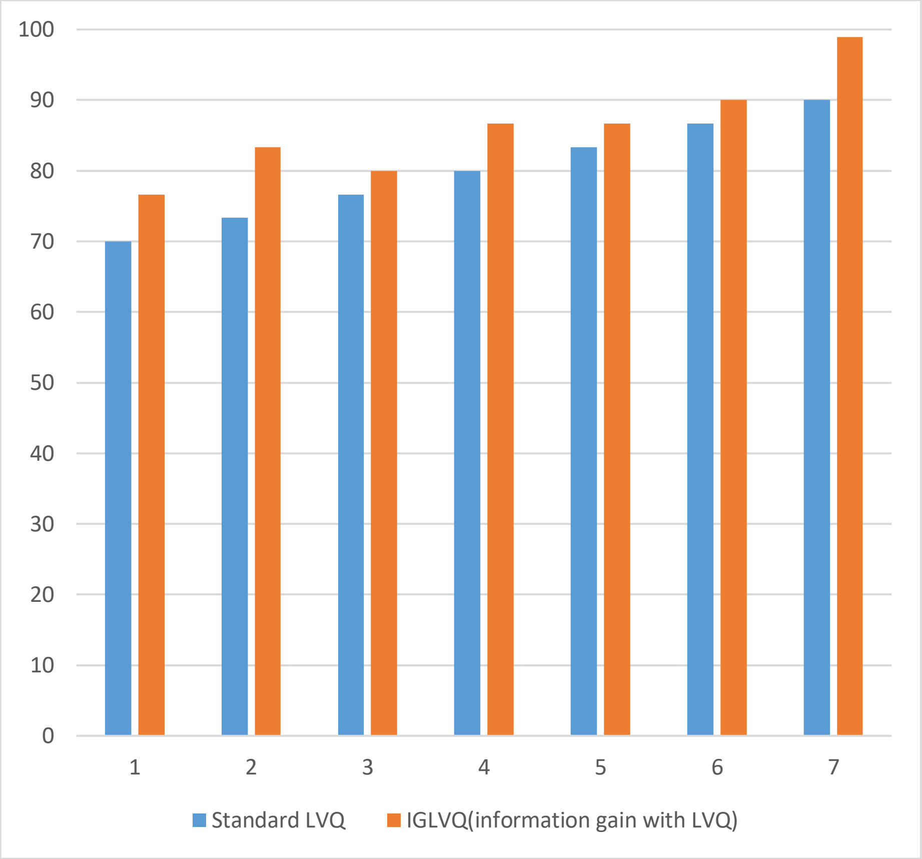


Fig 3. Comparison of LVQ and IGLVQ

Thus, the proposed approach of LVQ based on the information gain performs better as compared to its conventional approach. Further, L. Ali and et al has proposed a statistical model for this dataset considering all the attributes and achieved classification accuracy of 93.3%[12]. It is evident that the improvement in classification accuracy in the proposed method is significant.

Conclusions:

The IG based weighted LVQ serves the data preprocessing task before training which is the most important feature of deep learning neural network. The accuracy of the model increases as we remove redundant attributes and consider only those attributes which are instrumental in decision making. However, in our model, without feature selection a significant improvement in classification accuracy is achieved.

CONFLICTS OF INTEREST

There are no conflicts to declare.

REFERENCES

[1]. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554,

[2]. Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access*, *8*, 107562-107582.

[3]. Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. Computational and mathematical methods in medicine, 2017.

[4]. Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M. R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y., & Tayebi, R. M. (2019). New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health and Technology*, 1-12.

[5]. Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. Neural Computing and Applications, 25(3-4), 511-524.

[6]. Caliskan, A., & Yuksel, M. E. (2017). Classification of coronary artery disease data sets by using a deep neural network. The EuroBiotech Journal, 1(4), 271-277 ,

[7]. Ashraf, M., Rizvi, M. A., & Sharma, H. (2019). Improved Heart Disease Prediction Using Deep Neural Network. Asian Journal of Computer Science and Technology, 8(2), 49-54.

[8]. Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia computer science*, *120*, 588-593.

[9]. Poornima, V., & Gladis, D. (2018). A novel approach for diagnosing heart disease with hybrid classifier.

[10]. Malav, A., & Kadam, K. A. (2018). A hybrid approach for heart disease prediction using artificial neural network and K-means. Int J Pure Appl Math, 118(8), 103-110.

[11]. Alkhasawneh, M. S. (2019). Hybrid cascade forward neural network with elman neural network for disease prediction. Arabian Journal for Science and Engineering, 44(11), 9209-9220.

[12]. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}} $ Statistical Model and Optimally Configured Deep Neural Network. IEEE Access, 7, 34938-34945.

[13]. Miao, K. H., & Miao, J. H. (2018). Coronary heart disease diagnosis using deep neural networks. Int. J. Adv. Comput. Sci. Appl., 9(10), 1-8.

[14]. Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 304-319.

[15]. Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, *6*, 63279-63291.

[16]. Fazakis, N., Kostopoulos, G., Kotsiantis, S., & Mporas, I. (2020). Iterative Robust Semi-Supervised Missing Data Imputation. *IEEE Access*, *8*, 90555-90569,

[17]. Pratiwi, A. I. (2018). On the feature selection and classification based on information gain for document sentiment analysis. Applied Computational Intelligence and Soft Computing,

[18]. Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. Neural Computing and Applications, 25(3-4), 511-524.