Dimensionality Reduction of Healthcare Data through Niche Genetic Algorithm

Madhu H.K.
Assistant Professor
Department of MCA
BIT, Bengaluru, India

D. Ramesh, PhD
Professor& Head
Department of MCA
SSIT, Tumkur, India

ABSTRACT
Technology into medical health care has generated voluminous parameters for human physiological condition, forming data for high dimensions. Data which is raw makes any computing techniques complex and it is not structured.Structuring data can be a pre-processing model,but extracting useful parameters which contribute to reducing the computational complexities of any intelligent algorithm for classification and prediction is a big challenge in technology. Dimensionality reduction is a common strategy adopted by research to select appropriate parameters for further computations. In this research work Niche genetic algorithm is implemented on various healthcare datasets which extracts relevant parameters for classification and prediction of healthcare data with reduced computation complexity and increased accuracy. The proposed model is independent of any application, but restricts to structured data.

Keywords
Dimensionality reduction, Principal Component Analysis (PCA), UCI Health care dataset.

1. INTRODUCTION
Clinical decision system aid indiagnoses of diseases either for identification, classification and prediction. Clinical data is voluminous and high dimensional. It is a raw data extracted from various healthcare devices. Conversion of unstructured data to structured data format was a research topic decade ago. But today data cleaning, extracting relevant parameters [1,2,3], reconstructing missing data[4,5] are the major research as pre-processing models for intelligent decision systems. Most of the research work applied for feature extraction [6,7] and feature selection [6,7] as pre-processing models to considerrelevant data as input to the proposed classification and prediction model. Dimensionality reduction is one such approach where techniques reduce the number of input attributes in a data set and a detailed survey is discussed in section 2. In this research work feature selection & feature extraction are usedfor dimensionality reduction. Niche genetic algorithm is implemented with prior knowledge of niche radius or distance threshold and also self-adaptive mechanism for dimensionality reduction. The proposed method is discussed in detail in section 3. Section 4 highlights the results and discussion of the proposed work. Section 5 covers the conclusion of the proposed model.

2. RELATED WORK
The survey concentrates various dimensionality reduction techniques proposed by various research scholars considering feature extraction and feature selection [6,7]. A survey is also covered on Niche genetic algorithm [16,17,18,19] to understand the methodology implemented by various researchers.

G. Thippa Reddy et al. [1] proposed Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), two popular dimensionality reduction techniques, which are investigated on four popular Machine Learning algorithms DT, SVM, NBC and RFC. The authors used Cardiotocography dataset from the University of California repository. The findings of the experiments show that when the dimensionality of the datasets is high, ML methods with PCA generate better outcomes. When dataset's dimensionality is low, ML methods produce better results without dimensionality reduction. Beatriz Remeseiro et al. [2] reviewed recent approaches of feature selection in medical applications, demonstrating that the feature selection is a useful pre-processing tool that also reduces the number of input features, saving time, but also aids specialists in understanding the basic causes of certain diseases. The methods of dimensionality reduction approaches are discussed by Shaeela Ayesha et al. [3], as well as their appropriateness for various types of data and application domains. Furthermore, concerns with dimensionality reduction approaches have been brought out, which can have an impact on the accuracy and usefulness of the results.RazanAbdulhammed et al. [4] employs two methods for reducing feature dimensionality ie., Auto-Encoder and Principal Component Analysis. Authors used tocreate an IDS, the low-dimensional features obtained from both methodologies are utilised to construct various classifiers such as Bayesian Network, Random Forest, Quadratic Discriminant Analysis and Linear Discriminant Analysis. The CICIDS2017 dataset's attributes were reduced from eighty-oneto ten in this study and getting a high accuracy of 98.6 percent in multi-class and binary classification. Yashar Kiarashinejad et al. [5], show how to analyse, create, and optimise electromagneticnanostructures using a new computationally efficient approach based on deep learning (DL) approaches.Rizgar R. Zebari et al. [6] proposed that, the FS and FE methods were used to conduct dimensionality reduction. Because information is produced at an ever-increasing rate, FS is a useful way for reducing some severe dimensionality issues, such as reducing redundancy, deleting irrelevant data, and improving result comprehensibility.

To handle large-scale feature selection, Xubin Wang et al. [7], introduces SaWDE, a novel weighted meta-heuristic optimization method based on self-adaptive mechanisms. Authors present a new self-adaptive mechanism for capturing the varied properties of datasets from historical data by selecting several strategies from a strategy pool.Gurcan Yavuz and Dogan Aydın [8], proposed the "Self-adaptive Search Equation-based Artificial Bee Colony", is an ABC algorithm,

that may internally identify the optimal local search process and search equation during execution. The technique is a self-adaptive one that selects the best search equations for a given task by weeding out ineffective ones from a pool of randomly produced search equations.Qianqian Zhang et al. [9], proposed a further developed self versatile transformation calculation. Another versatile change administrator and versatile scaling factor in view of the self-versatile DE calculation are proposed to change the DE calculation's control boundaries and differential methodology. To build forecast execution, this work proposed by Shanshan Guo et al. [10], presents a novel multi-stage self-versatile classifier gathering model in view of measurable and AI procedures. To start, a multi-step information readiness strategy is utilized to change over the crude information into normalized information and produce more agent highlights. Second, in view of their exhibition in datasets, fundamental classifiers can be self-adaptively picked from the applicant classifier store, and their boundaries can be altered utilizing the Bayesian improvement process. Third, with these worked on base classifiers, the group framework is carried out, and it can deliver new elements through multi-facet stacking and get the classifier loads in the troupe model through molecule swarm streamlining. Kusum kumari Bharti and P.K. Singh [11], present three-stage aspect decrease models for eliminating pointless, excess, and loud elements from the first component space while preserving a significant amount of information. To construct a low dimension feature subspace, these models combine the benefits of the FS and FE approaches. The suggested three-stage models greatly increase the performance of the clustering method as assessed by, macro F-score, micro F-score, and total execution time in three text datasets with varied features.D.Napoleon and S.Pavalakodi[12], proposed Principal component analysis and linear transformation to reduce dimensionality and determine the initial centroid, which is then given to the K-Means clustering technique. Krzysztof siwek et al. [13] proposed linear and nonlinear reduction approaches and their efficiency will be compared. In this research, their application to the visualisation of various classes, as well as clustering and classification of data, will be investigated and explored.Rahmatwidia sembiring et al. [14] proposes a model for creating complex information bunching from wellbeing data sets is proposed. Particular Value Decomposition (SVD), Principal Component Analysis (PCA), Self-Organizing Map (SOM), and FastICA were utilized to diminish the quantity of aspects. The Improved Niche Genetic Algorithm (INGA) [15] is introduced by Min Zhu et al., in this study. In the structure of the specialty climate, it utilizes a self-versatile specialty separating activity to advance populace assortment and forestall nearby ideal arrangements. The INGA was approved in a sepsis patient characterization model. The outcomes uncover that utilizing INGA diminished the component dimensionality of datasets from 77 to 10 and that the model anticipated 28-day demise in sepsis patients with a precision of 92%, which is a lot higher than different procedures.

Basnamohammedsalihhasan and Adnan mohsinabdulazeez [16], reviews will begin with an introduction to the basic concepts of (PCA), followed by a description of some related topics and a discussion.Ersinkusetbodur and Donald douglas[17], objective of their study was to create and evaluate a filter algorithm for reducing the feature set in medical databases. The method binarizes the data, then analyses the risk ratio of each prediction with the reaction separately, producing ratios that describe the relationship between a predictor and a class feature. The given work provides the ideal value for the selection of relevant and non-redundant features for both text and micro-array datasets. Ten popular benchmark data were used to validate the performance of the proposed work by Divyajain and Vijendrasingh [18]. The exhibition and exactness of the most by and large utilized element extraction draws near, like EMD and PCA, just as component determination techniques, like connection, LDA, and forward choice, have been concentrated in this concentrate by [19] S.Velliangiri et al.

From survey, it is evident that Niche genetic algorithm [15] is efficient and best for classification and prediction. The available data is input into the proposed classification and prediction model where results show improved efficiency.

3. PROPOSED METHODOLOGY

NGA depends on the organic idea of a specialty being applied to transformative calculations. It portrays an endurance situation with a foreordained distance boundary L. The L of NGA is pre-set, permitting just a single uncommon person to contend in this distance.

INGA [15], is defined with Niche elimination process,

The Niche Elimination Procedure is as indicated in the part (I). Following that, INGA is built, as seen in part (II).

Part 1: Operation to eliminate Niche:

(a)The distance boundary *L*en is intended to be self-versatile with the Euclidean distance among people of every age to stay away from the union issue brought about by pre-set Len.

$$\mathrm{D} = ||Mi - Nj|| = \sqrt{\sum_{k=1}^{len}(mik - mjk)^{2,}} \quad \text{--- (i)}$$

$$i.j \in \{1,2,\ldots\ldots,X\}, i \neq j.$$

Mi and Nj are two members of the current population, both of whom are loci genetics. The present population's size is denoted by the letter M. The number of loci utilised to create and evaluate individual lengths is called len. The values of loci are m_{ik} and m_{jk}. The distance parameter Len is computed as follows:

$$L\text{en} = \min\{D\} \quad \text{--- (ii)}$$

Because each generation's individuals differ and the distance parameter's values change over time, an appropriate distance parameter will be found during each generation's evolution process in order to achieve a better niche habitat.

(b)Allowing only one exceptional person in Len will result in the elimination of other potentially excellent people who aren't as good as the one who is kept. As a result, the commonalities of biallelic loci are employed within the distance parameter Len to rate the similarity of people and determine whether they should be kept.

The two following equations that describe biallelic loci similarities and average similarities between the two individuals:

$$\mathrm{ED(Mi, Mj)} = \sum_{k=1}^{len} \frac{num\ (Mik == Mjk)}{len} \quad \text{---- (iii)}$$

$$i. \in \{1,2,\ldots\ldots,X\}, j \in \{i+1, i+2, \ldots, X\},$$

where ED(M_i, M_j) is the degree of similarity between two individuals, M_i and M_j, and num(M_{ik} == M_{jk}) denotes the number of people with the same allele value. Consider

$$\text{BEDi} = \frac{\sum_{j=i+1}^{M} ED(Mi,Mj)}{\text{len} * (X-1)} \quad \text{--- (iv)}$$

$$i. \in \{1,2, \ldots \ldots, X\}, j \in \{i+1, i+2, \ldots ., X\},$$

The mean comparability between the i^{th} individual and the others is addressed by BEDi. The comparability between two people will be recognized when {Mi-Mj} < Len. On the off chance that the closeness is more than the normal, the person with the lower wellness will be appointed a punishment work, as outlined in the situation beneath.

Individuals with a lesser level of fitness can be retained if necessary:

$$f_j^i(M) = f_j(m) * Q, \quad \text{--- (v)}$$

where $f_j(M)$ is the individual's previous fitness, f^1_j (M) is the current fitness, and Q is the penalty (often 10^{-30}). This strategy can help to lessen the number of people who are eliminated.

(c)The scale of a subpopulations should be regulated to sustain the population's diversity. As a result, (vi) and (vii) are built using a memory pool of ideal individuals to limit the scale for each generation's subpopulations:

$$\text{e(t)} = \sum_{i=1}^{X_{(t)}} \frac{f_i(t)}{X_{(t)}} \quad \text{----- (vi)}$$

where e(t) is the population's average fitness value, $f_i(t)$ is the fitness of individual I in generation t, and X(t) is the population's scale in generation t. As a result, in generation t+1, the scale of subpopulations is X(t+1). This is calculated using the formula

$$X_{(t+1)} = X_{(t)}.f(t).\frac{\text{t}}{\sum_{i+1}^{t} f(i)}. \quad \text{---- (vii)}$$

Excellent evolutionary individuals are exchanged through a memory pool of ideal individuals. The procedure raises the chances of obtaining more great individuals and, to some extent, eliminates the problem of premature convergence during a single population's evolutionary process. The general t+1 individuals are sorted by fitness, and the formers N are placed in the memory pool.

A memory pool of ideal people is intended to trade fantastic transformative people. The activity builds the chance of acquiring more astounding people, and somewhat, keeps away from the issue of untimely combination during the transformative course of a solitary populace. The people of general t+1 is arranged by wellness, and the formers N are placed into the memory pool.

Through the consequence of X(t+1), the capacity of keeping up with the populace variety, $d(p)$, is planned as in the accompanying two conditions. The more modest the worth of $d(P)$ is, the higher its populace variety is:

$$\text{d(q)} = \frac{\sum_1^t d(Q)_t}{\text{t}}, \quad \text{------- (viii)}$$

where $d(Q)_t$ is the generation t capability to preserve population diversity. And $d(Q)_t$ is constructed as follows:

$$\text{d}(Q)_t = \frac{1}{1.X_{(t)}} \sum_{j=1}^{1} max\left\{\sum_{i=1}^{X(t)}(1-a_{ij}), \sum_{i=1}^{n} a_{ij}\right\}, \text{ ---- (ix)}$$

where l is the length of the singular encoding, X(t) is the populace scale in age t, and a_{ij} is the i^{th} person's j^{th} locus.

Part 2: INGA:

Xstarting individuals are created at random at initially. According to completely reflect the benefit of controlling mistakes by integrating INGA with classifier, it is normal to utilize the complementary of the amount of error square of the classifier test set information as the fitness work [20].

$$\text{f(M)} = \frac{1}{\sum_{i=1}^{n}(r'_i - r_i)^2}, \quad \text{---- (x)}$$

where r' is the test set's projected value, r is the true value, and n is the test set's sample number. Individuals are arranged in descending order by fitness, and the first N are remembered in the memory pool (N<X).

To develop great initial people, a niche elimination operation is used. The fantastic initial individuals X(t) are created in this step.

Compute the Euclidian D between Mi and Mj using the formula (i). Second, determine the self-adaptive survival distance L using the formula (ii).

Criterion of similarity to evaluate if a person should be maintained, judge the similarity of the individuals within the distance L using the allele contrast approach. The frequency of biallelic loci and mean similarity between two people are examined when {Mi-Mj}<L is used. The individual with the lower fitness level does not have to be eliminated if they are not similar. The average similarity BEDi between two individuals is determined by (iii) and (iv). The similarity of biallelic loci ED(Mi, Mj). When ED(Mi, Mj) >BEDi, the punishment function fj(M) = fj(M) P is used to punish fj(M) according to (v). If not, the person with the lowest fitness level will be kept. When {Mi-Mj}>L, on the other side, the individual with the lowest fitness level is maintained.

As per (vii), the quantity of subpopulations X(t+1) is determined. People are arranged by wellness in diving request; if the size of the current subpopulation X(t) is more noteworthy than X(t+1), select the people X(t+1); in any case, N people are converged in the memory pool with the current subpopulations and arranged by wellness in dropping request; when N+X(t) > X(t+1), the previous people X(t+1) of (N + X(t)) are chosen; when N+X(t) < X(t+1). The original population will have a higher average fitness as a result of this strategy, which will be favourable to population evolution forward towards a solution to the problem.

When the probability of mutation and crossover is considered, it is either too small to keep the framework from falling into the nearby ideal optimal solution, or it is too large to prevent the system from falling into the local optimal solution but is prone to instability and convergence due to the high frequency of crossover and mutation. The equations of self-adaptive crossover (Pc) and mutation probability (Pm) are utilised to improve this shortcoming [21, 22]:

$$q_c = \begin{cases} Q_{c1} - \frac{Q_{c1}-Q_{c2}}{f_{max}-f_{avg}}\left(f^1 - f_{avg}\right) & f^1 \geq f_{avg} \\ Q_{c1} & f^1 < f_{avg}, \end{cases} \quad \text{- (xi)}$$

$$q_m = \begin{cases} Q_{m1} - \frac{Q_{m1}-Q_{m2}}{f_{max}-f_{avg}}\left(f - f_{avg}\right) & f \geq f_{avg} \\ Q_{m1} & f < f_{avg}. \end{cases} \quad \text{- (xii)}$$

The most extreme wellness esteem is fmax; the normal wellness worth of every populace is favg; the greater wellness

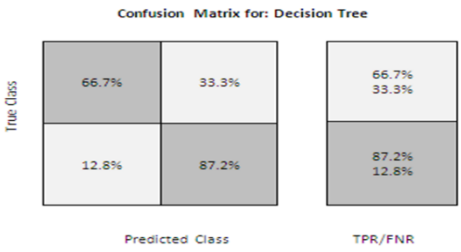
worth of two intersection people is f; and the wellness worth of change people is f. Pc1, Pc2 are the crossover and mutation probability values of two individuals, respectively; Pm1 and Pm2 are the mutation probability values of two individuals.

Put the new individual through the Niche elimination operation again after the self-adaptive mutation and crossover operation to get the best individual.

If it fails to achieve the termination criterion, increase the counter t to t+1 and change the populace to the new cutting-edge populace prior to continuing. If the termination condition is met, output the dimensionality reduction parameters that were chosen as the best.

4. RESULTS AND DISCUSSION

UCI heart disease dataset [23] consists of 76 attributes for 303 samples used to run the proposed methodology. Usually, many researchers consider 13 attributes.When these features were passed into INGA. Ten latent variables were extracted using cross validation, the extracted features entered into decision tree classification algorithm and confusion matrix evaluated.



It is shown that INGA extracted the features subset in a more appropriate control with less redundancy.

5. CONCLUSION

Health care data posse's challenges to technology as the data is unstructured and redundant. In this work dimensionality reduction technique is applied to select the subset of features for further classification and prediction. The INGA algorithms are a renovated technique from conventional genetic algorithm. This Niche creates a subsetfor extraction and selection. The appropriateness of feature selection depends on the accuracy achieved in classification and prediction.

6. REFERENCES

[1] G. Thippa Reddy, Praveen Kumar Reddy, KuruvaLakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava andThar Baker. "Analysis of Dimensionality Reduction Techniques on Big Data". https://creativecommons.org/lic. VOLUME 8, 2020. IEEE Access.

[2] Beatriz Remeseiroa and Veronica Bolon-Canedo. "A review of feature selection methods in medical applications". 2019 Elsevier Ltd. Computers in Biology and Medicine 112 (2019) 103375.

[3] Shaeela Ayesha, Muhammad Kashif Hanif and Ramzan Talib. "Overview and comparative study of dimensionality reduction techniques for high dimensional data". Information Fusion 59 (2020) 44–58. 2020 Elsevier.

[4] RazanAbdulhammed, Hassan Musafer, Ali Alessa and MiadFaezipour. "Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection". Electronics 2019, 8, 322; www.mdpi.com/journal/electronics.

[5] Yashar Kiarashinejad, Sajjad Abdollahramezani and Ali Adibi. "Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures". School of Electrical and Computer Engineering, Georgia Institute of Technology, 778 Atlantic Drive NW, Atlanta, GA 30332, USA.

[6] Rizgar R. Zebari, Adnan Mohsin Abdulazeez, Diyar QaderZeebaree, Dilovan Asaad Zebari and Jwan Najeeb Saeed. "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction". Journal of Applied Science and Technology Trends Vol. 01, No. 02, pp. 56 –70, (2020).

[7] Xubin Wang, Yunhe Wang, Ka-Chun Wong and Xiangtao Li. "A self-adaptive weighted differential evolution approach for large-scale feature selection". https://doi.org/10.1016/j.knosys.2021. 07633 0950-7051/© 2021 Elsevier.

[8] Gurcan Yavuz and Dogan Aydın. "Improved Self-adaptive Search Equation-based Artificial Bee Colony Algorithm with competitive local search strategy". https://doi.org/10.1016/j.swevo.2019.100582. 11 October 2019.

[9] Qianqian Zhang, Daqing Wang and Lifu Gao. "Research on the inverse kinematics of manipulator using an improved self-adaptive mutation differential evolution algorithm". International Journal of Advanced Robotic Systems. May-June 2021: 1–11 DOI: 10.1177/17298814211014413. journals.sagepub.com/home/arx.

[10] Shanshan Guo, Hongliang He and Xiaoling Huang. "A Multi-Stage Self-Adaptive Classifier Ensemble Model With Application in Credit Scoring". 2169-3536. 2019 IEEE. http://www.ieee.org/publications_standards/publications/rights/index.html.

[11] Kusum Kumari Bharti and P.K. Singh. "A three-stage unsupervised dimension reduction method for text clustering". Journal of Computational Science 5 (2014) 156–169. http://dx.doi.org/10.1016/j.jocs.2013.11.007.

[12] D.Napoleon and S.Pavalakodi . "A New Method for Dimensionality Reduction using KMeans Clustering Algorithm for High Dimensional Data Set". International Journal of Computer Applications (0975 – 8887). Volume 13– No.7, January 2011.

[13] Krzysztof Siwek, StanisławOsowski, Tomasz Markiewicz and Jacek korytkowski. "Analysis of medical data using dimensionality reduction techniques". PrzeglądElektrotechniczny, ISSN 0033-2097, R. 89 NR 2a/2013.

[14] Rahmat Widia Sembiring, Jasni Mohamad Zain and Abdullah Embong. "Dimension Reduction of Health Data Clustering". IJNCAA, 2011 (ISSN: 2220-9085).

[15] Min Zhu, Jing Xia, Molei Yan, Guolong Cai, Jing Yan and Gangmin Ning. "Dimensionality Reduction in Complex Medical Data: Improved Self-Adaptive Niche Genetic Algorithm". Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine.

Volume 2015, Article ID 794586, http://dx.doi.org/10.1155/2015/794586.

[16] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction". Journal Of Soft Computing and Data Mining. VOL. 2 NO. 1 (2021) 20-30.

[17] ErsinKusetBodur and Donald Douglas. "Filter Variable Selection Algorithm Using Risk Ratios for Dimensionality Reduction of Healthcare Data for Classification". Processes 2019, 7, 222; doi:10.3390/pr7040222. www.mdpi.com/journal/process.

[18] Divya Jain and Vijendra Singh. "An efficient hybrid feature selection model for dimensionality reduction". ICCIDS 2018. Procedia Computer Science 132 (2018) 333–341.

[19] S.Velliangiria, S.Alagumuthukrishnan and S IwinThankumar joseph. "A Review of Dimensionality Reduction Techniques for Efficient Computation". ICRTAC 2019.

[20] A. E. I. Brownlee, O. Regnier-Coudert, J. A. McCall, S. Massie, and S. Stulajter, "An application of a GA with Markov network surrogate to feature selection," International Journal of Systems Science, vol. 44, no. 11, pp. 2039–2056, 2013.

[21] C. W. Ho, K. H. Lee, and K. S. Leung, "A genetic algorithm based on mutation and crossover with adaptive probabilities," in Proceedings of the Congress on Evolutionary Computation (CEC '99), vol. 1, IEEE, Washington, DC, USA, July 1999.

[22] T. Ingu and H. Takagi, "Accelerating a GA convergence by fitting a single-peak function," in Proceedings of the IEEE International Fuzzy Systems Conference, vol. 3, pp. 1415–1420, Seoul, Republic of Korea, August 1999.

[23] UCI Heart disease data set. https://archive.ics.uci.edu/l/datasets/Heart+Disease.

7. AUTHOR'S PROFILE

Mr. Madhu H.K., is a research scholarat SSIT, Tumkur(SSAHE),having 21 years of Teaching experience at Department of M C A, in BIT, His research interest includes Data Miningand Big Data Analytics.

Dr. D. Ramesh, Professor and HOD from SSIT,Tumkur, India. His vision is to make SSIT as a centre ofexcellence for imparting technical knowledge in thefield of computer applications, nurturing technical competency and social responsibility among budding software professionals.