Optimized Auto Encoder on High Dimensional Big Data Reduction: an Analytical Approach

Arifa Shikalgar[1], Shefali Sonavane[2]
[1]Dept. of Computer Science & Engineering, Walchand College of Engineering Sangli,
Maharashtra, India – 416415
shikalgar.arifa@walchandsangli.ac.in
[1]Dept. of Information Technology, Walchand College of Engineering Sangli, Maharashtra, India - 416415

Abstract . Big data comprises of huge volume of data, which is exponentially increasing with time. Since the data is too large in size; the traditional data management tools are ineffective in processing these data effectively. The big data encompasses huge count of variables, hence analyzing each of the variables at a microscopic level is not feasible, as it might consume days or even months to have a meaningful analysis. This is time-consuming and costlier. Therefore, the Dimensionality Reduction (DR) techniques can be utilized. In general, the DR is a technique for reducing the count of input variables with fewer losses. These input features can cause deprived performance for ML algorithms. This paper introduces an optimized auto-encoder based dimensionality reduction model to deal with large datasets. The weight of the auto-encoder is fine-tuned by a selfadaptive Bumble Bees Mating Optimization (SA-BBMO) algorithm, which is the conceptual upgrading of standard BBMO. Further, to validate the appropriateness of the projected dimensionality reduction model, the experiments are conducted using big datasets. The corresponding results acquired are compared over the nonlinear dimensionality reduction techniques like PCA, K-PCA, LDA etc, in terms of Reconstruction error, Convergence, V-Measures, Silhouet Coefficient and Computation Time.

Keywords: Big Data; Dimensionality Reduction; Optimized Auto- Encoder; SA- BBMO

Nomenclature

reviation	cription
A	ptive Genetic Algorithm
SA-BBMO	-Adaptive Bumble Bees Mating Optimization
DA	-Linear Dimensionality Reduction
	ward Stepwise
GA	Adaptive Dimensionality Reduction Genetic Optimization Algorithm
CV	variance in the column values
	ensionality Reduction
	etic Algorithm
	correlation
	etic Algorithm
SA	rid Genetic And Simulated Annealing Algorithm
	ficial Bee Colony
CA	nel Principal Component Analysis
	fly Algorithm
	cle Swarm Optimization
A	ar Discriminant Analysis
DRM	Hybrid Dimension Reduction Method

	uped Feature Extraction
V	o of missing values
C	lidates and split columns
	er Score
	vard feature construction
	om forest
	rmation Gain
CA	se Locality For Principal Component Analysis
	ure selection
	al network
	cipal Component Analysis
	ward feature elimination
	ral Network

1 Introduction

Big data is a huge volume of dataset with multi-level labels and diverse sets of information [9][10][11] [12]. It encompasses volume of information, speed or velocity at which the data is collected or created, variety of the data points; therefore it is often referred as “big data’s three V's". Here, the volume is the most important aspect. In general, the big data is categorized as: “unstructured or structured”. Structured data encompasses an extremely structured format in which the data can indeed be "retrieved, processed and stored" in a rigid format. However, the unstructured data do not have a structure or specific form [6] [7] [8]. Therefore, it is highly complex and tedious to analyze and process these unstructured data. Further, there are multiple benefits in processing Big Data like: cost –efficient, time saving, superior operational effectiveness, Early recognition of risk to the product/services, Improved customer service as well [13] [14] [15]. Further, big data is quietly significant in diverse applications like Healthcare for disease prediction, Academia for allround development of budding learners via digital courses, Banking for fraud detection, Manufacturing for improving the supply strategies and product quality, and Transportation for route planning, traffic control, road congestion management. As the technology is revalorizing, the data being generated is expanding, and these data could not be handled by traditional database systems. So, it is good to employ the data-dimensionality reduction techniques in the datasets that is encompassed with massive volume of data columns [16] [17] [18].

In general, the DR is a new technique to diminish the quantity of input variables. The dimension reduction is more often utilized for solving machine learning problem by selecting the better features for classification as well as regression. Moreover, the statistical and machine reasoning methods face diverse issues like hardware orientation, higher computational complexity, while dealing with these high dimensional data [19][20][21]. The seven most commonly used techniques for data-DR techniques are: “RMV, LV-CV, HC between two columns, PCA, C-SC in a RF, BFE and FFC”. These techniques are commonly utilized for reducing the dimensions of the data, by removing columns that higher information or add no new information. This might increase the computational complexity and the processing time. Therefore, the deep learning models can be utilized to diminish the dimensionality of data. But, here the time taken to train the deep learning model is longer. So, the optimization concept [22 ~ 30] can be utilized.

The major contribution of this research work is: Introduces an optimized auto-encoder based dimensionality reduction model, where the weight of the auto-encoder is fine-tuned using a self-adaptive Bumble Bees Mating Optimization (SA-BBMO) algorithm, which will be conceptual improvement of standard BBMO.

The remaining part of this paper is arranged as: Section II addresses the recent works in DR, Section III tells about the proposed novel optimized auto-encoder based dimensionality reduction model using self-adaptive bumble bees mating optimization. Then, Section IV talks about the attained results. This paper is completed in Section V.

2 Literature review

2.1 Related works

In 2020, Fong *et al*. [1] have proposed a novel NL-DA technique based on the bottleneck deep autoencoders. In this research work, two-fold contributions were introduced: (a) In the bottleneck deep autoencoders, the monotonicity constraint was introduced for determining the single nonlinear component; (b) The multiple nonlinear components were estimated using the proposed FS deep learning architecture. The proposed work was tested using two real data, and the resultant had exhibited better results in terms of reconstruction errors.

In 2020, Kuang *et al*. [2] have introduced an ADRGA with the intention of overcoming the problem of dimensionality in the big data. Here, when the adjacent dimension angle of the individual data is less than the angle factor, then the dimension of the data is considered to be smaller and it is marked as 0. Further, the proposed ADRGA model was tested with "eight high-dimensional test functions". The proposed work was better than existing techniques like AGA, standard GA and HGSA with respect to "convergence, accuracy, and speed". In 2020, Li *et al*. [3] have developed SLPCA with the intention of resolving the DR issues in the big data. In addition, they have introduced a R2P-PCA with the objective of considering the trade-off in between the performance like efficiency and complexity. The outcomes have demonstrated that the proposed SLPCA model was consistent than the extant techniques in terms data reconstruction error and clustering accuracy. In 2020, Li *et al*. [4] have developed a novel FH-DRM by integrating the GFE and the multi-strategy feature selection for removing both the redundant and the irrelevant information. Initially, the authors have used the the maximum likelihood estimation method to set the intrinsic dimensionality of original data. Then, the irrelevant features were removed using the multi-strategy methods like the FS and IG based FS". Further, the redundant information in every cluster was removed using the PCA based feature extraction. The proposed method had exhibited excellent efficiency.

3 Proposed novel optimized auto-encoder based dimensionality reduction model using self-adaptive Bumble Bees Mating Optimization

3.1 Overall Description

In this research work a novel optimized auto-encoder based dimensionality reduction model on large datasets is presented. In general, Autoencoder is an unique NN with three layers. The weight of the layers is generally adjusted with certain training methods; to narrow the result generated to being closer to the input values given. In this work, the weight of the auto-encoder is fine-tuned using a SA-BBMO algorithm, which is conceptual improvement of standard BBMO.

3.2 Optimized Auto- Encoder

An Autoencoders is an special unsupervised three-layer NN structure with "output layer, hidden layer and input layer" within itself. In addition, the Autoencoder has an decoder as well as an encoder. The input data $\left(Data^{input}\right)$is mapped onto the hidden layer by the encoder, and the newly acquired features ae shown in Eq. (1). $l = f(m) = k\left(Wt_{(1)}l + p_{(1)}\right)$ (1)

In which, the input vector and output vector is $m \square Q^{h \square 1}$ and $l \square Q^{s \square 1}$, respectively. The weight function in the hidden layer is $Wt^{(i)} \square Q^{s \square h}$ and input bias is $bi^{(i)} \square Q^{s \square 1}$. In addition, the notation h,s,k represents the dimension of the input data, count of hidden layer units and activation function, correspondingly. Moreover, the de-coder takes the responsibility of transforming the remapped data in the hidden layer to the original data with reduced features. Mathematically, the decoding mechanism is shown in Eq. (2). $m = q(l) = k\left(Wt_{(2)} l + p_{(2)}\right)$ (2)

In addition, the reconstruction error R_{error} and the cost function D are mathematically manifested in Eq. (3) and Eq. (4), respectively.

$$R_{error} = \left\| m - q(f(m)) \right\|_2 \quad (3)$$

$$D(Wt,bi) = \square\square N1 \square_{v^N=1} (12 \overline{m_{(v)}} - \overline{q\left(f\left(m_{(v)}\right)\right)^2) \square\square\square} + \square$$

$$+ \square 2_{o \square 2=1} \square_{vk-o1kw} \square_{o=+1i} \left(Wt_{wv(o)}\right)_2 \quad (4)$$

In which, the notations $m^{(v)}$, $Wt_{wv}{}^{(o)}$ depicts the v^{th} sample, the v^{th} unitof o^{th} layer and w^{th} unitof $(o+1)^{th}$ layer, respectively. Further, the count of samples is denoted as N , and in the o^{th}layer, the count of samples is denoted as k_o .

In addition, the deionising of auto encoders is done to reduce the noisy data. Mathematically, the reconstruction error of the denoising $R_{d(error)}$ is shown in Eq. (5). The mathematical expression for reconstruction error in denoising D_{den} is shown in Eq. (6).

$$R_{d(error)} = \left\| m - q(f(m_1)) \right\|_2 \quad (5)$$

$$D_{den}(Wt,bi) = \square\square N1 \square_{vN=1} (12 \overline{m_{o\,(ov+)1}} - \overline{q\left(f\left(m_{1(v)}\right)\right)^2) \square\square\square} + \square$$

$$+ \square\square_2 \square_{k\,k} \square_{(o)} \Big)_2$$

$$\left(Wt_{wv}\right.$$

$$2_{o=1v-1w=1}$$

Here, the notation m ,m_1 ,l and x represents the initial input data, corrupted input data, the new feature obtained by the corrupted input data and the output obtained by decoding the acquired new feature. The major intention behind this research work is to reduce thereconstruction error in denoising D_{den} during the dimensionality reduction process. Thereby, the weightWt is fine-tuned by SA-BBMO. Mathematically, the objective function for the proposed algorithm is given as in Eq. (7). The architecture of auto encoder is illustrated in Fig.1.

$$Obj = \min(D_{den}) \quad (7)$$

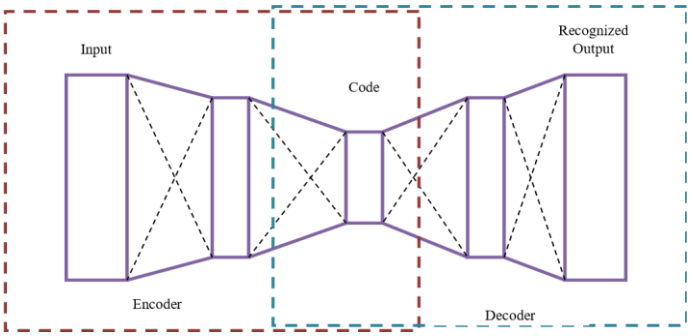


Fig. 1. Architecture of Auto-encoder

3.3 SA-BBMO

The standard BBMO was introduced with the stimulation acquired from the bumble bee's mating behavior. Three different categorizes of mating behavior are utilized in theBBMO algorithm, they are : "the queen, the workers and the drones (males)". The BBMO is a renowned optimization model, with a major advantage of providing the global solutions without getting trapped into the local optima. But, here the convergence seems to be lower. Therefore, a SA-BBMO is established. The steps followed in the self-adaptive Bumble Bees Mating Optimization are depicted below:

Step 1: The initial population $(Pop)$of the search agents (initial population) is generated. In addition, the maximal iteration is denoted as Max^{iter} , and the current iteration is *iter* .

Step 2: The fitness function is computed for every bumble bee using Eq. (7)

Step 3: The search agent with the most appropriate fitness is selected as the bumble bee

Step 4: The rest of the bees are considered as the drones **Step 5:**Sort the drones based on the fitness.

Step 6: For matching, the queen selects the drones, and then the queen of the drones' are stored to spermatheca of the queen.

iter

Step 7:: While $Max \square iter$ do

Step 8: Perform crossover operator for generating the broods. Here, select two crossover operator number $(Cross_1, Cross_2)$, which is the parameter for controlling the parameters of the queen and drones. This $Cross_1$ is generated randomly within the limit 0,1(i.e. $random[0,1]$). If $random[0,1] \square (Cross_1)$and $random[0,1] \square (Cross_2)$, then inherit the corresponding value from the queen, else select the corresponding value from a drones' genotypes. Thus, the solution of i^{th} brood is $brood_{ij}(iter)$; *j*Dimension of the problem , and $queen_j(iter)$

represents the solution of the queen, and $dist_{kj}(iter)$is the solution of k^{th} drone. Dimension of the problem

$$\square queen_j(iter), if\ random[0,1] \square (Cross_1)$$

$$brood_{ij}(iter) = \begin{cases} Worker_r(iter), & if\ random[0,1] < (Cross_2) \quad (8) \\ dist_{kj}(iter), & Otherwise \end{cases}$$

Further, the value of $Cross_1$ and $Cross_2$ are computed using Eq. (9) and Eq. (10), respectively. These $Cross_1$ and $Cross_2$ controls the range of values.

$$Cross_1 = (Upper_{bound} - Lower_{bound}) * * W_1 - \frac{MaxW_{1iter} * iter}{} + Lower_{bound} \quad (9)$$

$$Cross_2 = (Upper_{bound} - Lower_{bound}) * * W_2 - \frac{MaxW_{2iter} * iter}{} + Lower_{bound} \quad (10)$$

Here, W_2 and W_1 are the parameters

Then, for each brood compute the fitness using Eq. (7).

Step 9: The broods get sorted on the basis of computed fitness '.The brood with the most appropriate fitness is selected as the new queen

Step 10: The rest broods are the workers

Step 11: The workers as well as the old queens feed the new queens

Step 12: The genotypes of the genotypes are mutated to create the percentage of the drones

Step 13: The genotypes of the workers' are mutated for creating the rest of the drones **Step 14:**Then, for each drone compute the fitness using Eq. (7).

Step 15: The drones that move away from the hive are determined based on the Lévy flights using Eq. (11).

$$dist_{i,j} = dist_{i,j} + (dist_{k,j} - dist_{i,j}) + Levy() * random + \frac{}{LevyFlightmotion} \quad (11)$$

In which, $dist_{i,j}$, $dist_{k,j}$ and $dist_{i,j}$ denotes the solutions of *i*, *j*,*k* drones, respectively. In addition, represents the parameter, which is utilized for computing the parameter of drone *i* that is affected by *k*,*l* drones.

Step 16: Based on the computed fitness, the drones are sorted.

Step 17: While for each new queen, when the utmost amount of matings is not reached, do

Step 18: Then, for mating, each of the queen chooses the drones

Step 19: The genotypes of the drones' are stored to each spermatheca of the spermatheca

Step 20: Then, for next iteration, the new queens survival is found

Step 21: Expiry of drones and workers

Step 22: ReturnThe best queen as the most favourable solution

4 Results and discussions

4.1 Simulation procedure

The proposed novel auto-encoder based dimensionality reduction model was implemented in MATLAB and it was tested using the big data collected from:

Genome datasets(Dataset1): This dataset is collected from : "https://ftp.ncbi.nlm.nih.gov/genomes/refseq/" [Access Date: 2012-01-24]. It is available for all "eukaryotes through the NCBI Datasets Genomes web interface and include genome, transcript and protein sequence, annotation and a data report".

BBC News Classification (Dataset2): This dataset is collected from: "https://www.kaggle.com/c/learn-ai-bbc" [Access Date: 2012-01-24]

Heart Disease Data Set(Dataset3): It is collected from :" https://archive.ics.uci.edu/ml/datasets/heart+disease" [Access Date: 2012-01-24]. It encompasses Seventy six attributes.

Breast Cancer Prediction Dataset(Dataset4): It is downloaded from: "https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset" [Access Date: 2012-01-24]. The corresponding results acquired is compared over the extant nonlinear dimensionality reduction techniques like PCA, K-PCA, LDA , and optimization Algorithms like GA+ auto-encoder, FF+ auto-encoder, ABC+ autoencoder and PSO+ auto-encoder in terms of Reconstruction error, Convergence, VMeasures, Silhouette Coefficient and Computation Time.

4.2 Reconstruction Error Analysis

The proposed tactic (SA-BBMO+ auto-encoder) for dimensionality reduction is compared over the existing models in terms of Reconstruction error to be calculated using Eq.(5) and (6). Since, the objective behind the current research work is to lessen the Reconstruction error, the technique with the lowest Reconstruction error will be suggested as the best technique for dimensionality reduction. The Reconstruction error results acquired for Dataset1, Dataset2, Dataset3 and Dataset4 are shown in Fig. 2. On have a view on the produced graphical results, the Reconstruction error of SA-BBMO+ auto-encoder seems to be equivalent to the existing techniques like GA+ auto-encoder, FF+ auto-encoder, ABC+ auto-encoder, PSO+ auto-encoder, PCA, K-PCA and LDA, respectively. But, while analysing the acquired results mathematically, the SA-BBMO+ auto-encoder is found to be superior to the existing techniques with less error. On observing the Reconstruction error of SA-BBMO+ auto-encoder and existing techniques for dataset-1, the SA-BBMO+ auto-encoder is 22.1%, 14.2%, 18.9%, 11.7%, 57.1% , 92.5% and 85% better than the extant tactics like the GA+ auto-encoder, FF+ autoencoder, ABC+ auto-encoder, PSO+ auto-encoder, PCA, K-PCA and LDA, respectively. Similar to this, the SA-BBMO+ auto-encoder exhibits better results than the existing techniques for all other datasets taken into consideration. Therefore, the proposed work is suggested as an apt technique for dimensionality reduction.

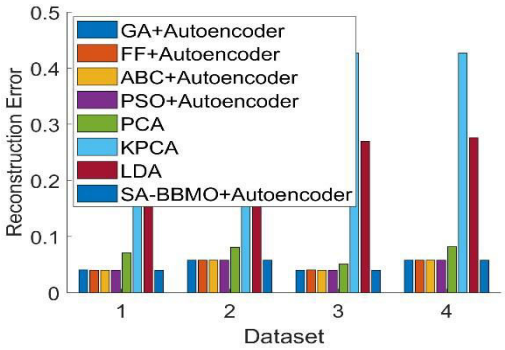
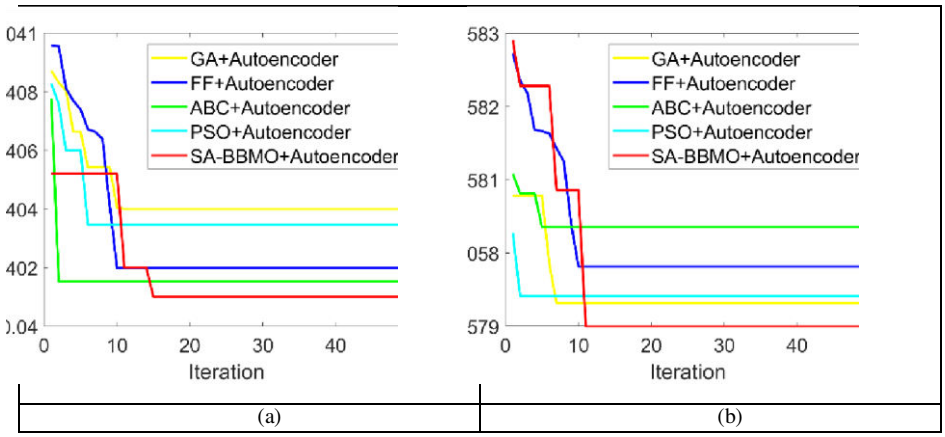
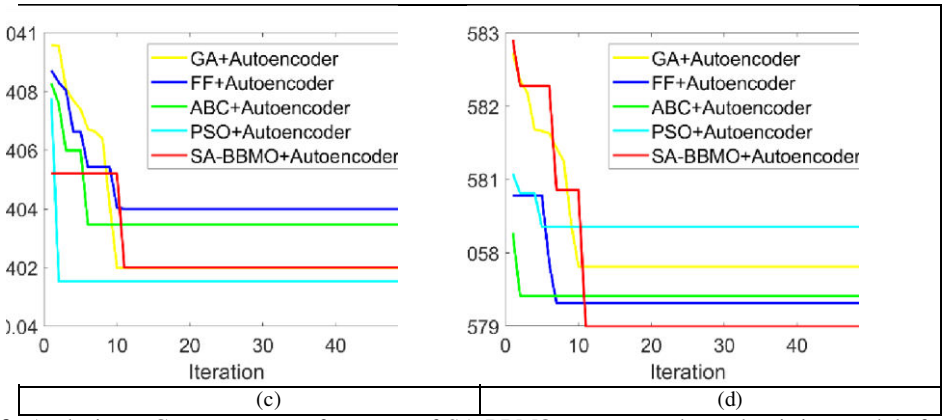


Fig. 2. Analysis on Reconstruction error performance of SA-BBMO+ auto-encoder and existing models for dataset1, dataset2, dataset3 and dataset4

4.3 Convergence Analysis

The convergence analysis strongly portrays about the achievement of the described objective or fitness function. In this research work, the major objective is reduction in the reconstruction error. The approach with the least cost function is said to have achieved the specified objective. The results acquired in terms of convergence analysis for dataset1, dataset2, dataset3 and dataset4 is manifested in Fig. 3 (a), Fig. 3(b), Fig. 3(c) and Fig. 3(d), respectively. This assessment is accomplished by altering the iteration counts. On analysing the acquired results, the cost function of both the proposed tactic (SA-BBMO+ auto-encoder) as well as existing tactics seems to be higher. However, as the count of iterations tends to increase, there is a gradual fall in the cost function. Even though, the cost function of the proposed as well as existing model had reduced, a best performance (i.e. least cost function) is recorded with the proposed work in all the collected datasets. On observing the acquired cost function of datset-1, the least cost function is recorded by the proposed work at the maximal count of iteration (i.e.50th iteration). Therefore, at the 50th iteration, the SA-BBMO+ auto-encoder is 0.24%, 0.4%, 0.74% and 0.99%improved over the existing models like ABC+ autoencoder, FF+ auto-encoder, PSO+ auto-encoder and GA+ auto-encoder, respectively. In a similar way, the SA-BBMO+ auto-encoder shows the least cost function for higher count of iterations. As a whole, the proposed work is said to have achieved the specified objective is a robust manner.





(c) (d)

Fig. 3. Analysis on Convergence performance of SA-BBMO+ auto-encoder and existing models for (a)dataset1, (b)dataset2, (c)dataset3 and (d)dataset4

4.4 Analysis on V-Measures

In general, the Validity Measure (V-measure) is a "Validity Measure (V-measure) "that tells about the success of a technique in keeping the ground truth labels as the baseline [30]. Mathematically, the V-measure can be computed using the weighted harmonic mean of *H* and *C* using Eq. (10).

$$\text{Validity Measure}_{\square} = 1 - \frac{(1+\square).H.C}{\square.H + C} \quad (12)$$

In which, *H* denotes the homogeneity and *C* denotes the completeness. The resultant of the Validity Measure acquired for both the existing and proposed tactic is shown in Table I. For datset-1, the computed Validity Measure of SA-BBMO+ auto-encoder is 34.6%, 5.6% and 26.7% better than the existing techniques like PCA, K-PCA and LDA, respectively. Similarly, the achieved Validity Measure of SA-BBMO+ autoencoder for dataset-2 is 0.60617, which if the highest value, and it is 70.8%, 63% and 70% better than existing techniques like PCA, K-PCA and LDA, respectively. In all the other collected datasets also, the proposed work records the highest Validity Measure.Therefore, the proposed work is said to be much appropriate for dimensionality reduction.

set	PCA	CA	LDA	SA-BBMO+ a
set1				
set2				
set3				
set4		75		

Table 1. Analysis on Validity Measure of SA-BBMO+ auto-encoder and existing technique

4.5 Analysis on Silhouet Coefficient

It is a measure of distance, and it tells about the closeness of the data point in a cluster [30]. The value of the silhouette coefficient ranges between -1 and 1. In case, if the silhouette coefficient is 0, then the particular data point is said to be between 2 clusters inflection point. The mathematical formula for silhouette coefficient is expressed in Eq. (13).

$$\text{silhouette coefficient(I)} = 1 - \frac{Dis(I) - Dis^{avg}(I)}{\max[Dis(I), Dis_{avg}(I)]} \quad (13)$$

Here, $dist(I)$ denotes the I^{th} data point's mean distance from all other points re-

$Dist^{avg}(I)$ represents the I^{th} data siding in the group, at which it exists. In addition,
point's smallest average distance from all points in the group, at which it exists. On analyzing the acquired results in terms of silhouette coefficient, the proposed work exhibits higher performance for all the datasets. For datset3, the acquired silhouette coefficient is 18.7%, 37.2% and 74.7% better than the existing techniques like PCA, KPCA and LDA, respectively. Similar to this, the SA-BBMO+ auto-encoder achieves the highest value for all the datasets, and it is evident from the results shown in Table II. Therefore, the proposed work is recommended for solving the dimensionality reduction issues in big data.

set	PCA	CA	LDA	SA-BBMO+ au

Table 2. Analysis on silhouette coefficient of SA-BBMO+ auto-encoder and existing technique

4.6 Analysis on Computation Time

The required computational time for processing each of the datasets for efficient dimensionality reduction is manifested in Table III. The proposed work records the least computational time as 63.4 seconds in dataset2 and datse4, respectively.

set	PCA		LDA	SA-BBMO+ au
set1			68	
set2			31	
set3			15	
set4			52	

Table 3. Analysis on computational time of SA-BBMO+ auto-encoder and existing technique

4.7 Reduction Vs Loss of Quality

In these days, the dimensionality reduction plays an attractive role, owing to the expansion in the volume of data generated each day. In general, the dimensionality reduction technique keeps on reducing the initial features considerably, until a set of permit features are generated for the original properties of the data. Nevertheless, this reduction entails an inherent loss of quality, which affects the consideration of data, during the analysis.

This section portrays about the errors recorded by the proposed work after employing the proposed dimensionality reduction model. Here, the error (loss of quality) is computed by varying different reduction variations. For dataset1, the original dimension of data is taken as 15873 and the by dimensionality reduction model, the varied reduced dataset values as 7937, 5291, 3968, and 3175 for

each iteration. Therefore, the error value is recorded as 0.0027, 0.0028766, 0.0028766 and 0.0037736. The results acquired are tabulated in Table IV. As per the recorded results, the dimensionality of the data is concentrated to a greater extends, after the application of the proposed dimensionality reduction model. But the most interesting thing is, the errors recorded during this dimensionality reduction are negligible. Therefore, the loss in relevant data is also insignificant, and by this means the loss of quality is lower. For this reason, the proposed work is suggested as an apt technique for dimensionality reduction in big datasets.

		ed	Error		ed	Error		ed	Error		ed	Error
						766			766			736
			526			526			526			526
	6	3	263			53			53			833
	14	7							32			12

Table 4. Reduction Vs Loss of Quality of Proposed Work for , Dataset2, Dataset3 and Dataset4

5 Conclusion

In this research work, a novel optimized auto-encoder based dimensionality reduction model was developed to deal with large datasets. Then, the weight of the autoencoder was fine-tuned by a SA-BBMO algorithm, which is the conceptual improvement of standard BBMO. Finally, the corresponding results acquired were compared over the existing techniques in terms of Reconstruction error, Convergence, VMeasures, Silhouet Coefficient and Computation Time. The achieved Validity Measure of SA-BBMO+ auto-encoder for dataset-2 is 0.60617, which if the highest value, and it is 70.8%, 63% and 70% better than existing techniques like PCA, PCA and LDA, respectively.

References

[1] Youyi Fong, Jun Xu,"Forward Stepwise Deep Autoencoder-based Monotone Nonlinear Dimensionality Reduction Methods", Journal of Computational and Graphical Statistics, 2020

[2] Tai Kuang, Zhongyi Hu , and Minghai Xu,"A Genetic Optimization Algorithm Based on Adaptive Dimensionality Reduction", Mathematical Problems in Engineering , 2020

[3] Pei Heng Li, Taeho Lee, and Hee Yong Youn,"Dimensionality Reduction with Sparse Locality for Principal Component Analysis", Mathematical Problems in Engineering, 2020

[4] Mengmeng Li , Haofeng Wang , Lifang Yang , You Liang ,Zhigang Shang , Hong Wan,"Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction", Expert Systems With Applications, 2020

[5] R. M. Gahar, O. Arfaoui, M. S. Hidri and N. B. Hadj-Alouane, "A Distributed Approach for HighDimensionality Heterogeneous Data Reduction," IEEE *Access*, vol. 7, pp. 151006-151022, 2019. doi: 10.1109/ACCESS.2019.2945889

[6] D. Kaur, G. S. Aujla, N. Kumar, A. Y. Zomaya, C. Perera and R. Ranjan, "Tensor-Based Big Data Management Scheme for Dimensionality Reduction Problem in Smart Grid Systems: SDN Perspective," IEEE *Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1985-1998, 1Oct.2018.doi: 10.1109/TKDE.2018.2809747

[7] L. Kuang, L. T. Yang, J. Chen, F. Hao and C. Luo, "A Holistic Approach for Distributed Dimensionality Reduction of Big Data," IEEE *Transactions on Cloud Computing*, vol. 6, no. 2, pp. 506-518, 1 AprilJune2018.doi: 10.1109/TCC.2015.2449855

[8] M. A. Da Silva Lopes, A. D. Dória Neto and A. De Medeiros Martins, "Parallel t-SNE Applied to Data Visualization in Smart Cities," IEEE *Access*, vol. 8, pp. 11482-11490, 2020. doi: 10.1109/ACCESS.2020.2964413

[9] X. Wang, W. Wang, L. T. Yang, S. Liao, D. Yin and M. J. Deen, "A Distributed HOSVD Method With Its Incremental Computation for Big Data in Cyber-Physical-Social Systems," IEEE*Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 481-492, June 2018. doi: 10.1109/TCSS.2018.2813320

[10] L. Kuang, L. T. Yang and Y. Liao, "An Integration Framework on Cloud for Cyber-Physical-Social Systems Big Data," IEEE *Transactions on Cloud Computing*, vol. 8, no. 2, pp. 363-374, 1 April-June 2020.doi: 10.1109/TCC.2015.2511766

[11] G. T. Reddy *et al*., "Analysis of Dimensionality Reduction Techniques on Big Data," IEEE *Access*, vol. 8,pp.54776-54788,2020.doi: 10.1109/ACCESS.2020.2980942

[12] D. Hong, N. Yokoya, J. Chanussot, J. Xu and X. X. Zhu, "Joint and Progressive Subspace Analysis (JPSA) With Spatial-Spectral Manifold Alignment for Semisupervised Hyperspectral Dimensionality Reduction,"IEEE*TransactionsonCybernetics*.doi: 10.1109/TCYB.2020.3028931

[13] K. Wang, C. Xu, Y. Zhang, S. Guo and A. Y. Zomaya, "Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid," IEEE *Transactions on Big Data*, vol. 5, no. 1, pp. 34-45, 1 March 2019. doi: 10.1109/TBDATA.2017.2723563

[14] P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar and M. Palaniswami, "A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data," IEEE *Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 641-654, 1 April 2019. doi: 10.1109/TKDE.2018.2842191

[15] A. Zare, A. Ozdemir, M. A. Iwen and S. Aviyente, "Extension of PCA to Higher Order Data Structures: An Introduction to Tensors, Tensor Decompositions, and Tensor PCA," in *Proceedings of the IEEE*, vol.106,no.8,pp.1341-1358,Aug.2018.doi: 10.1109/JPROC.2018.2848209

[16] S. Ramírez-Gallego *et al*., "An Information Theory-Based Feature Selection Framework for Big Data Under Apache Spark," IEEE *Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 9, pp.1441-1453,Sept.2018.doi: 10.1109/TSMC.2017.2670926

[17] M. Mohajeri, A. Ghassemi and T. A. Gulliver, "Fast Big Data Analytics for Smart Meter Data," IEEE *Open Journal of the Communications Society*, vol. 1, pp. 1864-1871, 2020. doi: 10.1109/OJCOMS.2020.3038590

[18] X. Zhou, Y. Hu, W. Liang, J. Ma and Q. Jin, "Variational LSTM Enhanced Anomaly Detection for Industrial Big Data," IEEE *Transactions on Industrial Informatics*. doi: 10.1109/TII.2020.3022432

[19] R. Hang and Q. Liu, "Dimensionality Reduction of Hyperspectral Image Using Spatial Regularized Local Graph Discriminant Embedding," IEEE *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3262-3271, Sept. 2018. doi: 10.1109/JSTARS.2018.2847042

[20] M. Yamada *et al*., "Ultra High-Dimensional Nonlinear Feature Selection for Big Biological Data," IEEE*Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1352-1365, 1 July 2018. doi: 10.1109/TKDE.2018.2789451

[21] Rajakumar B R, "Optimization using lion algorithm: a biological inspiration from lion's social behavior", Evolutionary Intelligence, Special Issue on Nature inspired algorithms for high performance computing in computer vision, Vol. 11, No. 1-2, pages 31-52, 2018, DOI: https://doi.org/10.1007/s12065-018-0168-y.

[22] B. R. Rajakumar, "Impact of Static and Adaptive Mutation Techniques on Genetic Algorithm", International Journal of Hybrid Intelligent Systems, Vol. 10, No. 1, pages: 11-22, 2013, DOI: 10.3233/HIS-120161

[23] Aloysius George, B. R. Rajakumar and Binu Dennis, "Genetic Algorithm based airlines booking terminal open/ close decision system", In proceedings of International Conference on Advances in Computing, Communications and Informatics, pages: 174-179, August 3-5, Chennai, India, 2012, DOI: 10.1145/2345396.2345426

[24] B. R. Rajakumar, "Lion algorithm and its Applications", Frontier Applications of Nature Inspired Computation in Springer Tracts in Nature-Inspired Computing (STNIC), Springer, Editors: Mahdi Khosravy, Neeraj Gupta, Nilesh Patel, Tomonobu Senju

[25] W.Brajula and M.Bibin Prasad, "ODFF Opposition and Dimension based Firefly for Optimal Reactive Power Dispatch", Journal of Computational Mechanics, Power System and Control, Vol.1,No.1, pp.110,2018.

[26] Amolkumar Narayan Jadhav,Gomathi N, "DIGWO: Hybridization of Dragonfly Algorithm with Improved Grey Wolf Optimization Algorithm for Data Clustering", Multimedia Research, Vol.2,No.3, pp.1-11,2019.

[27] kulkarni,Senthil Murugan T, "Hybrid Weed-Particle Swarm Optimization Algorithm and C- Mixture for Data Publishing", Multimedia Research, Vol.2,No.3, pp.33-42,2019."

[28] Jyothi Mandala and Dr. M.V.P. Chandra SekharaRao, "HDAPSO: Enhanced Privacy Preservation for Health Care Data", Journal of Networking and Communication Systems, Vol.2,No.2, pp.10-19,2019.

[29] G.K. Shailaja,Dr C.V. Guru Rao., "Impact of Opposition Intensity on Improved Cuckoo Search Algorithm for Privacy Preservation of Data", Journal of Networking and Communication Systems, Vol.2,No.4, pp.33-41,2019.

[30] Uditha Maduranga, Kalana Wijegunarathna, Sadeep Weerasinghe, Indika Perera and Anuradha Wickramarachchi, "Dimensionality Reduction for Cluster Identification in Metagenomics using Autoencoders", 2020 international conference on ICT for emerging region (ICTer), 2020