Successful Data Mining: With Dimension Reduction

Sariya Begum Syed Meraj[1](✉) and B. S. Shetty[2]

[1] Departement of Information Technology, S.G.G.S. Institute of Technology Nanded, Nanded, Maharashtra, India
syeduzma778@gmail.com

[2] Department of Information Technology, S.G.G.S. Institute of Technology Nanded, H.O.D, Nanded, Maharashtra, India
bsshetty@sggs.ac.in

Abstract. The technique of collecting important knowledge characteristics from a dataset in order to further transform it into usable information is known as data mining. With the combined use of statistics and machine learning, data mining has grown in popularity for a variety of applications, including better decision-making, revenue and operation optimization, cost reduction, anomaly detection, and many more. Despite recognising patterns, One of the most challenging jobs in machine learning is clustering. It is difficult to build an acceptable number of clusters because doing so could decrease the effectiveness of training and assessment. In engineering and scientific applications, the clustering value has steadily increased during the past few years. Many clustering techniques have low classification accuracy, and if we utilise a lot of data with larger dimensions, it will affect the performance and required storage of the algorithms. To reduce this we have to reduce the diamensions of dataset so that clustering algorithm performance will increases and required space will decreases so in this approach we are going to reduce the dimensions of dataset on AFRBFNN [23] algorithm. It is predicated on RL ideas. It categorises every pattern that the two techniques discussed earlier were unable to categorize [23]. Its misclassification rate has decreased. When compared to the other techniques, this model delivers the best classification accuracy. This approach is also quick and has less overlapping. So that we added PCA (Principal Component Analysis) to reduce the dimensions of the dataset and comparing the results with [23] before applying PCA and after applying PCA.

Keywords: Radial basis function neural network · Fuzzy membership function · Fuzzy clustering Fuzzy set HyperSphere · Principal Component Analysis

1 Introduction

Understanding numerous events in daily life requires the use of data analysis. It is quite challenging to comprehend. It takes a lot of time and has a lot of data. Consequently, in order to save time and make things easier to better understand, we examined a PCA method that can do several dimensions is reduced to two dimensions. The PCA approach

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 11–22, 2023.
https://doi.org/10.2991/978-94-6463-136-4_3

compresses the most data. Information is included in the matrix's first two columns, referred to as the main elements. We are going to implement this on one [12] clustering algorithm. Cluster analysis is a straightforward examination that needs little to no prior knowledge of research done in many societies. The notion of increasing intra-class equality and reducing inter-class equality is used to organise objects. Because of this, clustering, also known as exploratory data analysis, is widely used to build clusters from labelled or unlabeled data. For various data sets, a number of clustering algorithms have been created, and it has been discovered that the outcomes vary depending on the algorithm. Due to its unsupervised nature, clustering is the most difficult topic in machine learning. To avoid the issues that cause training and testing efficiency to drop, the right number of groups is being developed. Reduce the amount of time needed to create and retrieve groups. There is independence between the clustering technique and the data being used. In recent years, clustering has become increasingly significant in engineering and scientific applications. Over the past ten years, clustering analysis has received a lot of attention in publications. To solve scientific applications in the real world, several researchers have put forth and employed various strategies. Nearly all clustering techniques have flaws and work with some types of data. As a result, there is a continuous need for research into various clustering, but before doing that if we reduce the dimension's of that particular dataset then algorithm performance will improve and system performance also increases.

1.1 Review of Literature

In machine learning, neural networks are frequently employed for pattern classification. Over the past three decades, a lot of studies have been done. The problems of pattern classification, non-linear data sets are mapped to linear data sets in various ways. Researchers also employs fuzzy sets and the radial basis function. Simpson (1993) [4] proposed using a hybrid neuro-fuzzy system with a hyperbox to do supervised classification. It employs a three-step learning method. Hyperbox expansion, contraction, and overlap tests are the three types of tests. U. V. Kulkarni and T. R. Sontakke introduced the fuzzy hypersphere neural network (FHSNN) in 2001. The model is designed to recognise handwritten characters. U. V. Kulkarni et al. 2002 [5] expanded on fuzzy hypersphere neural network research and developed a new model, the General Fuzzy Hypersphere Neural Network (GFHSNN). U. V. Kulkarni et al. [6] suggest supervised learning using fuzzy set hyper line. He divided the dataset into hyper lines and classified it. Online training, rapid learning, and soft decision- making are also encouraged. Clustering was proposed by U. V. Kulkarni et al. [7] using a hyper line and a fuzzy set. Unsupervised learning is employed in this case. Online training is available, and recall is extremely fast. Classes are represented by fuzzy set hyper line segments. According to Anas Quteishat and Chee Peng Lim, a repurposed fuzzy minmax neural network with rule extraction features uses a Euclidean distance metric for prediction, a rule extraction approach, and pruning. [8]. A. B. Kulkarni et al. [3] expanded prior research work in 2018 and built a new Fuzzy Hypersphere Neural Network (FHNN) classifier. The RBF notion has been incorporated into the learning model. In 2014 According to Sehgal S, Singh H, Agarwal M, Bhasker V One of the methods used for dimension reduction is Principal Component Analysis (PCA) pattern recognition methods, among other uses, is used to analyses challenging high-dimensional data simply by taking a glance at the vast amount of information. After performing some data analysis, we must reduce the high dimension of the data by constructing a low dimension layout and then interpreting the results [11].

1.2 Dimensionality Reduction

Applications for data mining in bioinformatics, risk management, forensics, and other fields require very high-dimensional datasets. The "Curse of Dimensionality," a well-known issue, is brought on by the abundance of dimensions. Due to the inclusion of several meaningless and irrelevant dimensions or features in the dataset, this issue causes machine learning classifiers to perform less accurately. Finding the critical dimensions for a dataset that drastically decreases the number of dimensions involves using a variety of approaches. These approaches to feature reduction and subset selection result in a smaller feature set, which ultimately leads to higher classification accuracy and less expensive machine learning algorithms [13]. Principal Components Analysis (PCA), which identifies linear combinations of the variables in a training data set that maximise the variance explained by each linear combination, is the most widely used technique for determining a lower dimensional representation of a data set. we might wish to display the data if you're working on a large research project with plenty of data and examples that are each characterized by a variety of characteristics. Therefore, you must decrease the dimension and put your data in a location with a dimension that is no larger than 3D. Principal Component Analysis can be used to do this (PCA). PCA may also be used to accelerate the machine learning algorithm's training phase. Machine learning algorithm becomes more complicated and requires learning more parameters as you add more features. Consequently, the computation time will take longer.

Figure 1 shows the flow diagram for proposed work.

In the Fig. 1 we mentioned the working flow of our proposed work. In which few steps are there:

- Step 1: Here in this step we collect the dataset it may be raw dataset so we used one of the popular UCI dataset i.e. Iris dataset for multiple dimentions.
- Step 2: loading the dataset.
- Step 3: Standardize data; without data standardisation, PCA will not be able to identify the best principal components since PCA is very sensitive to datasets with large levels of variation in their values. Here we are using StandardScaler() function in python for implementing this.

$$\text{X\_scale} = \text{scaler.transform(X)}$$

- Step 4: Applying PCA with components 2, We are using the PCA function of Sklearn. Decomposition module. Then we will get the principal components with 2d. Then we will get 2D principal components.
- Step 5: Applying PCA with principal components of 3 it means for 3D then we will get 3D principal components.

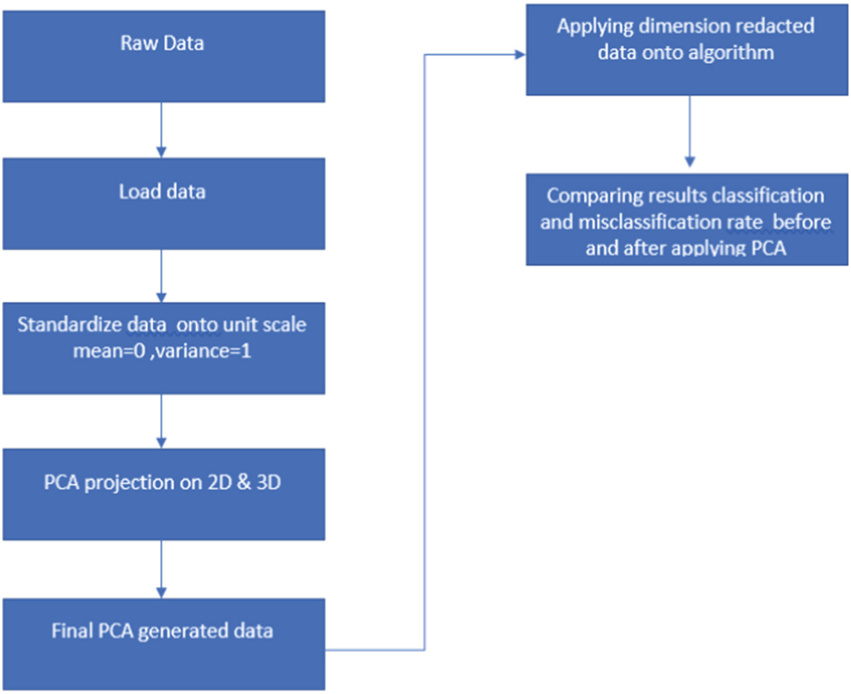

Raw Data
Load data
Standardize data onto unit scale mean=0 ,variance=1
PCA projection on 2D & 3D
Final PCA generated data
Applying dimension redacted data onto algorithm
Comparing results classification and misclassification rate before and after applying PCA

Fig. 1. Flow diagram of proposed system

- Step 6: Applying PCA 2D generated data into one of the Machine learning algorithms [14] first without applying PCA then with applying PCA.
- Step 7: then capture the classification and misclassification rate before and after applying PCA and compare them.

2 Learning Algorithm

2.1 StandardScaler Method

CommonScaler Method Rescaling one or more characteristics to have a mean value of 0 and a standard deviation of 1 is the process of data normalization. This technique is more affective if dataset attribute has distribution, is Gaussian. After scaling data, we have predicted model and seen that accuracy is best comparing to another method. xi – mean (x)/ stdev(x). When a dataset is standardized, the value distribution is scaled to have a mean of 0 and a standard deviation of 1. The Sklearn model selection function train-test splits data arrays into training data and testing data subsets. This function eliminates the requirement for manual dataset division. By default, Sklearn train test split will divide the two subsets into random divisions.

Each feature or variable is scaled to unit variance once the mean has been removed using StandardScaler. This process is carried out independently based on features. Since

StandardScaler estimates the empirical mean and standard deviations of each feature, it might be affected by outliers (if they are present in the dataset).

Transformer: Classes called transformers make it possible to transform data while also preparing it for machine learning. Simple Imputer, MinMaxScaler, Ordinal Encoder, and Power Transformer are just a few Scikit-Learn transformer examples. We occasionally need to conduct data transformations that are not built into popular Python libraries.

2.2 PCA Components and Variance

The amount of variation in a dataset that can be ascribed to each of the principal components (eigenvectors) produced by the principal component analysis (PCA) approach is measured statistically as "explained variance." It simply refers to how much of a dataset's variability may be attributed to each unique primary component. In other words, it reveals how much of the overall variation each component explains to us. This is significant because it enables us to priorities the most significant components when evaluating the findings of our investigation and to rank the components in order of significance. Let's take the example of creating a machine learning model to forecast home prices. The explained variance would indicate the extent to which the model can account for the variation in home prices. A bigger explained variance in this situation would be preferable since it would indicate that the model is more accurate in predicting house prices. When determining the relative importance of each component, the explained variance idea is helpful. Generally speaking, a major component's influence is more significant the more variance it can explain. PCA is a method for reducing the number of dimensions in data. This is accomplished by locating the directions in which the data vary most widely and projecting the data onto those directions. The "explained variance" is the amount of variance that each direction can account for. A reduced dataset's number of dimensions can be chosen using explained variance. It may also be applied to judge a machine learning model's quality. A model with high explained variance will often have strong predictive ability, whereas a model with low explained variance would not be as reliable.Let's use examples to better grasp the idea of explained variance. For instance, if we were to use PCA to reduce a dataset with 100 samples and 10 features to two dimensions, we would anticipate that the first component would account for around 86% of the variance (9/10) and the second component would account for approximately 14% (1/10). Various PCA models may also be compared using explained variance. For instance, we would claim that the latter model is better at describing the data if we compared two models that both reduce a dataset from 10 dimensions to 2, but one model only explains 80% of the variation while the other only explains 95% of the variance.

Algorithm reducing dimensions by PCA and applying it into learning model.

- Step 1: Collecting dataset (it may be raw dataset)
- Step 2: Loading dataset
- Step 3: Standardization of dataset
- Step 4: Applying PCA on ND dataset with principal components = 2(2D)
- Step 5: Applying PCA on ND dataset with principal components = 3(3D)
- Step 6: Applying PCA resulting data 2D, 3D one by one on OSFC algorithm [24].

- Step 7: Comparing the results of classification and misclassification rate before applying PCA and after Applying PCA.

3 Existing Concept on 2D Dataset

Example In this experiment, 2-dimensional examples with 2 classes and 15 samples are used to better understand the algorithm [24] and perform necessary classifications. The sample and its class labels are given in Table 1. Here 2 more unlabeled patterns (17.5, 8) and (15, 20) are taken for classification.

After successfully compiling the algorithm [24], it creates two groups for a given class. The nucleus of the Class 1 cluster (7.5, 7.5) has a radius of 3.5, as shown in Fig. 2.

The Class 2 nucleus (14, 14) has a radius of 2.9, as shown in Fig. 3.

Example 1-Distance matrix

Figure 4 shows groups formed after training all patterns by algorithm [24]. A detailed description of this cluster structure using the MSFC algorithm is given in [3]. We are given 2 new patterns with features values (6, 6) and (16.5, 16.5). These patterns are labelled, and our model has to classify its correct class. As per AFRBFNN approach, these patterns are unclassified, and our model cannot predict the label of such patterns.

Table 1. Case Study 1-Example Data Set

Sr. No.	Pattern Class	Total Count	Feature Vectors
1	1	8	(5, 5), (6.25, 5.5), (6.2,6.5), (7.25,7.5), (8.25,8.5), (9.25, 8.5), (5.2, 9.5), (10, 10)
2	2	7	(12, 12), (12.5, 11.5), (11.5, 13), (14, 14), (15, 15), (16, 16), (17, 17)
3	Unknown	2	(17.5, 8), (15, 20)

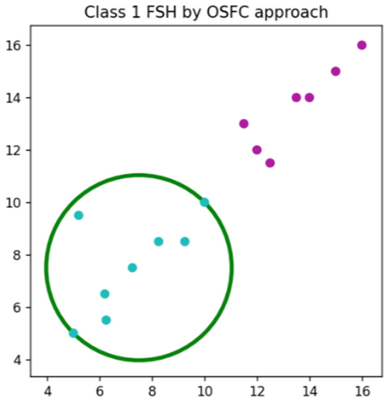


Fig. 2. Class 1 FSH by OSFC approach

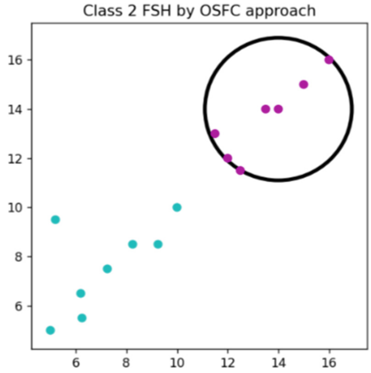


Fig. 3. Class 2 FSH by OSFC approach

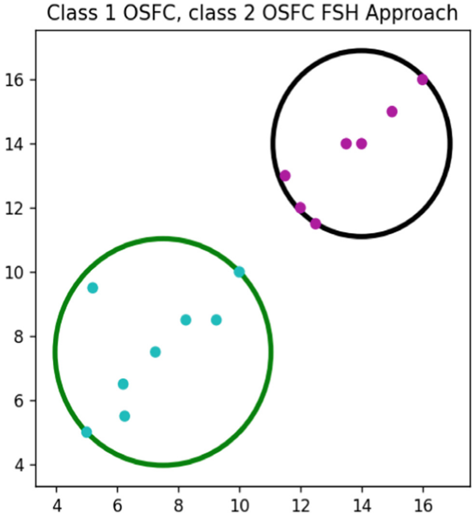


Fig. 4. Class 1 OSFC, class 2 OSFC

In this approach, we use the hybrid approach. We take existing model's intelligence and apply reinforcement approach which gives us the label of these patterns. Here model has collection of hyperspheres which consist of set of centroid and radius. We calculate the Euclidean Distance from these patterns to all cluster's centroid. The centroid which gives the minimum value, we assign the respective centroid's class to that pattern. So let us follow this simple step now. As per Reinforcement Spread Fuzzy Clustering (RSFC) algorithm of RAFRBFNN approach given by B.S. Shetty, we Calculate Euclidean Distance for unidentified pattern to all hyperspheres. We have collection of all hypersphere vector of all classes in HS[i] vector where i denotes hypersphere number. Let us determine the class for pattern (6, 6). Algorithm calculates the Euclidean Distance between (6, 6).

Table 2. UCI IRIS Dataset Description

Dataset	Data size	Features	Classes
Iris	150	4	3

FSH Approach and centroid of all hyperspheres of all classes and save it in distance vector. Then it selects the least distance value and class of respective hypersphere is assigned as the class of current pattern (6, 6). As per the algorithm [24], Table 2 shows the distance matrix from every input testing patterns to the cluster. Here Distance between pattern P1 and cluster of class-1 is 2.12 which is less than the distance between pattern P1 and cluster of class-2 which is 15.6 Hence class of cluster-1 is assigned to P1 which is class-1. Let us check for pattern (16.5, 16.5). Algorithm calculates the Euclidean Distance between (16.5, 16.5) and centroid of every class and save it in distance vector. Then it selects the least distance value and class of respective hypersphere is assigned as the class of current pattern (6, 6.0). As per the algorithm, we get distance = 5. 59017. Then We calculate with another class.

In the same way, we calculate the class of Pattern P2. Here, distance between pattern P2 and cluster of class-2 is 3.8 which is less than the distance between pattern P1 and cluster of class-1 which is 6.67. Hence class of cluster-2 is assigned to P2 which is class-2.

Now on this Approach [24] we have to reduce the dimensions of input(dataset) and apply PCA concept on that Algorithm [24]. For this we can't take this 2D dataset because its already in lower level 2D Dimensions so we are taking multiple dimensions data for UCI i.e., Iris Dataset and will apply our proposed algorithm and compare the results before applying PCA and after Applying PCA.and we will check the classification rate whether it will remain same or it will reduce.

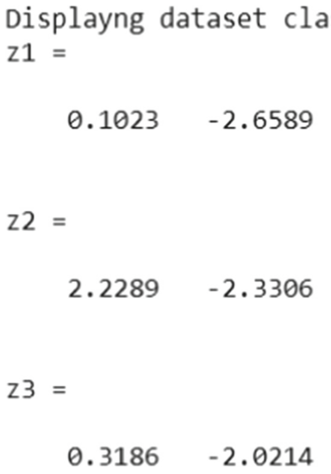
4 Case Study

In case study, the performance of the algorithm [24] is evaluated on IRIS dataset. The dataset contains a set of 150 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class (Species).

Before applying PCA on dataset the misclassification rate is 0 and classification rate is 100 percent for other dataset According to [24], but we are using that concept on Iris dataset with data size 150 and number of classes 3, so after implementation the results are same 100% classification rate. Now we are going to apply the dimensionality reduction technique on dataset by PCA with number of principal components equals 2 then we will check that after reducing dimensions it will decrease or remains same, Following is the class wise results first on 2 components and second one is for 3 components.

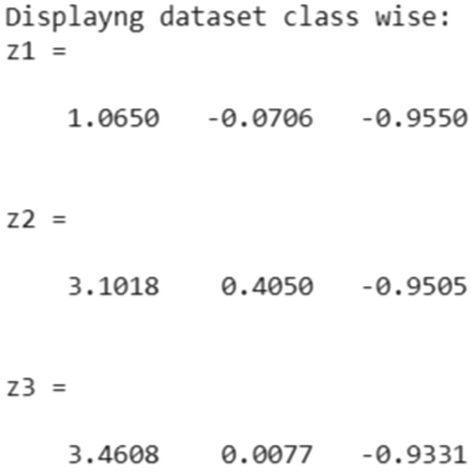
Samples of Class 1, Class 2 and Class 3 are shown in blue, red and green, respectively, and is shown in Fig. 5.

In Fig. 5 and Fig. 6 we can see that we have done for both principal components 2 as well as 3 we are getting accurate results that we can see below, hence by using Table 4 we can say that even after reducing the dimensions by 2D and 3D the results are same,



```
Displayng dataset class wise:
z1 =

    0.1023   -2.6589


z2 =

    2.2289   -2.3306


z3 =

    0.3186   -2.0214
```

Fig. 5. Applying PCA for components = 2 on Iris dataset according to proposed algorithm



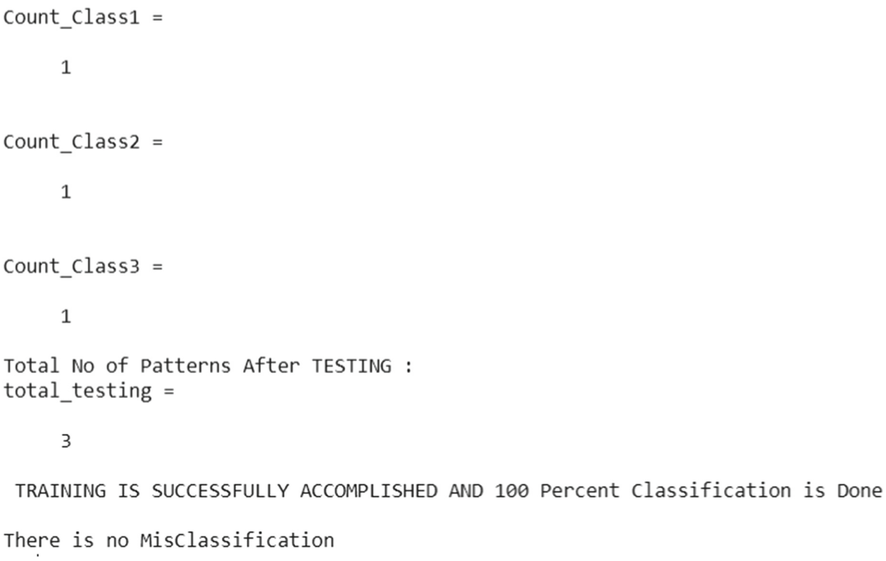
```
Displayng dataset class wise:
z1 =

    1.0650   -0.0706   -0.9550


z2 =

    3.1018    0.4050   -0.9505


z3 =

    3.4608    0.0077   -0.9331
```

Fig. 6. Applying PCA for components = 3 on Iris dataset according to proposed algorithm

Table 3. Classification Rate Before Applying PCA on dataset for components = 2 & 3

Dataset	Classification	Misclassification
Iris	100%	0(NO)

so by remains the same results we can decrease the time complexity and required space so the performance of algorithm will increase and system performance also increases (Fig. 7 and Table 3).



```
Count_Class1 =

     1


Count_Class2 =

     1


Count_Class3 =

     1

Total No of Patterns After TESTING :
total_testing =

     3

 TRAINING IS SUCCESSFULLY ACCOMPLISHED AND 100 Percent Classification is Done

There is no MisClassification
```

Fig. 7. Final result after applying PCA on iris dataset according to proposed algorithm

Table 4. Comparing results for iris dataset

Dataset	By using OSFC Classification rate	By reducing dimensions Classification rate for 2 components	By reducing dimensions Classification rate for 3 components	Misclassification rate for all
Iris	100%	100%	100%	0

5 Conclusions

- Any dataset may receive 100% training accuracy thanks to the training model's architecture.
- Because it uses an iterative process, the proposed RSFC method increases training accuracy by new patterns are always recognized and even though it reduces the dimensions it gives 100% classification rate.
- Its iterative methodology increases the rate of pattern identification and decreases the rate of misclassification even though we reduce the dimensions.
- It reduces the time complexity and space required for dataset.
- It speeds up OSFC deep learning Algorithm.
- It becomes easier to visualize the data when reduced to very low dimensions such as 2D and 3D.
- It becomes easier to visualize the data when reduced to very low dimensions such as 2D and 3D.

References

1. Mr. B.S. Shetty, Dr. U.V. Kulkarni, Ms. Preetee, M. Sonule, Ms. Pratiksha Patni.: Fuzzy Hypersphere Neural Network with Ant Colony Optimization related Rule Extraction. Journal of Applied Science and Computations (5), (2018). 10.10089. JASC.2018.V5I12.453459.1500167.
2. Balaji S Shetty Manisha S Mahindrakar and U.V. Kulkarni.: Advance Fuzzy Radial Basis Function Neural Network Algorithm. AISC Series Springer (Scopus indexed). International conference on Computing in Engineering Technology, Jan (2021).
3. A.B. Kulkarni, S.V. Bonde, U.V. Kulkarni.: A Novel Fuzzy Clustering Algorithm for Radial Basis Function Neural Network. International Journal on Future Revolution in Computer Science and Communication Engineering, (4),751—756(2012). 10.10089.JASC.2018.V5I12.453459.1500167.
4. P.K. Simpson.: Fuzzy min-max neural networks Classification. In IEEE Transactions on Neural Networks, 3(11), 769—783 (2000). https://doi.org/10.1109/72.159066.
5. Kulkarni Uday V., and T.R. Sontakke.: Fuzzy hypersphere neural network classifier. In 10th IEEE International Conference on Fuzzy Systems. (Cat. No. 01CH37297), vol.3(2001).
6. J.M. Waghmare and U.V. Kulkarni.: Unbounded Recurrent Fuzzy Min-Max Neural Network for Pattern Classification. 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1--8,(2019). https://doi.org/10.1109/IJCNN.2019.8852310.
7. Patil, Pradeep M and Dhabe, PS and Kulkarni, Uday V and Sontakke, TR.: Recognition of handwritten characters using modified fuzzy hyper line segment neural network. The 12th IEEE International Conference on Fuzzy Systems, pp. 1418-- 1422, (2003).
8. Quteishat, Anas and Lim, Chee Peng.: A modified fuzzy min–max neural network with rule extraction and its application to fault detection and classification. Applied Soft Computing, vol. 8, no.2, pp. 985--995, Elsevier, (2008).
9. Naeem, Muddasar and Rizvi, Syed Tahir Hussain, and Coronato.: A gentle introduction to reinforcement learning and its application in different fields. vol. 8, pp. 209320--209344, IEEE, (2020).
10. Frank A.: UCI Machine Learning Repository. Irvine, CA. University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml, vol. 216, pp. 261--267, (2016).
11. Sehgal S, Singh H, Agarwal M, Bhasker V.: Data analysis using principal component analysis. In 2014 international conference on medical imaging m-health and emerging communication systems (MedCom) . pp. 45--48, IEEE, (2014).
12. Shetty BS, Mahindrakar MS, Kulkarni UV.: Advance Fuzzy Radial Basis Function Neural Network. In: Applied Information Processing Systems. pp. 11--24, Springer, (2022).
13. N. Sharma and K. Saroha.: Study of dimension reduction methodologies in data mining. International Conference on Computing, Communication & Automation. pp. 133--137(2015), doi: https://doi.org/10.1109/CCAA.2015.7148359.
14. Shetty, B.S., Mahindrakar, M.S., Kulkarni, U.V., “Unbounded Fuzzy Radial Basis Function Neural Network Classifier. In: Iyer, B., Ghosh, D., Balas, V.E. (eds) Applied Information Processing Systems,” Advances in Intelligent Systems and Computing, vol 1354, Springer, https://doi.org/10.1007/978-981-16-2008-9_3