ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Optimal and Fast Health Data Clustering Using Hybrid Meta Heuristic Algorithm

V. Shanu[1], S.Vydehi[2]

M.Phil Scholar, Dept. of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India[1]

Professor & Head Of the Department, Dept. of Computer Science , Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: Several health oriented studies used machine learning approaches for health data analysis to detect health risks from different attributes of patient health records. Diagnosis and detection of diseases from patient electronic health records are complicated due to its instable nature and achieving that is a very promising area of research. From the numerous health data, the proposed system handles two popular disease dataset such as liver and heart related diseases. The proposed system performs effective clusters with high accuracy on high dimensional health data set. Even though there are several approaches developed to improve the cluster accuracy in the literature, but still some issues arises while performing the high dimensional dataset. So, the system proposes a new hybrid approach, which concentrates on the effective feature selection and data clustering. The proposed system is aimed to perform Effective pre-processing, feature selection and clustering. The pre-processing stage eliminates inter and intra cluster related problems. The second stage is the feature selection process, which performed using WPCA (Weighted Principle Component Analysis) and effective clustering using IBAT (Improved BAT Algorithm). The system implements a new Meta heuristic algorithm with the use of effective weighted features from the PCA. The system developed with the intension of high accuracy and less clustering time.

.

KEYWORDS: Medical Data Mining, Meta Heuristic Algorithm, PCA, IBAT Algorithm, Machine Learning, Data Mining, Clustering**.**

I. INTRODUCTION

Data mining is an integration of multiple disciplines such as statistics, machine learning, neural networks and pattern recognition. It is concerned with the process of computationally extracting hidden knowledge structures represented in models and patterns from large data repositories. Healthcare is a data intensive process. Many processes run simultaneously producing new data every second. It is a research intensive field and the largest consumer of public funds. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no longer ignore these emerging tools. This has resulted in unification of healthcare and computing to form Health care. They typically work through an analysis of medical data and a knowledge base of clinical expertise and it is an emerging field. In [1] authors described the need and algorithms of data mining in healthcare, in medical areas today, data collection about different diseases as very important. Medical and health areas are among the most important sections in industrial societies. The extraction of knowledge from a massive volume of data related to diseases and medical records using the data mining process can lead to identifying the laws governing the creation, the development of epidemic diseases [2][3]. Clustering is a predominant component such medical data mining process.

There are huge number of methods are developed to cluster. However, the existing clustering has many drawbacks such as being attentive in local optima, as well as local maxima and being sensitive to initial cluster centres'. One method to improve the clustering performance is hybridizing it with efficient optimization methods. In this paper, we use the WPCA algorithm to find optimal weighted feature among many and this will improve clustering accuracy. Proposed method experimental results compared with existing PCA and Firefly based algorithms. The results

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

show that the proposed algorithm has a higher efficacy than the other algorithms. To make the disease analysis in an effective manner, the data should be pre-processed, conditioned, and clustered. The study aims to improve the cluster performance and cluster quality through utilizing the WPCA with effective Meta heuristic algorithms. And the results of WPCA will be applied as a weighted feature for the next iteration into the MBAT algorithm. This iterative result improves the clustering efficiency and reduces the time for feature selection.

II. PROBLEM DEFINITION

Discovery of new information in terms of patterns or rules from large amounts of data is based on the machine learning technique [4]. Medical data mining, analysis and prediction play an important role in data mining. Diagnosis of a disease requires the performance of a number of tests on the patient. However, use of data mining techniques, can reduce the number of tests. This reduced test set plays an important role in time and performance. Liver data mining is important because it allows doctors to see which features or attributes are more important for diagnosis such as age, weight, etc. This will help the doctors diagnose liver more efficiently. There are various data mining techniques in use in healthcare industries but the research that has to be done is on the performance of the various clustering techniques, to enable the choice of the best among them can be chosen.

Table 1.0 meta-heuristic algorithm comparison table

Algorithm	Firefly	Cuckoo Search	Bat algorithm	Krill Herd
Year	2008	2009	2010	2012
Develop by	X.S Yang	X.S Yang, Suash Deb	X.S Yang	Gandomi and Alavi
Based on	Flashing behavior of fire fly	Obligate brood parasitism of cuckoo	Echo location behavior of micro bat	Herding behavior of krill
Objective function defined by	Brightness(light intensity) and attractiveness	Colour of eggs	Pulse rate emission and velocity	Distance from the food source
Features	High convergences rate, robust rate. Finds good optimum solutions in less number of iterations.	Implementation is simpler	Accurate and efficient	Efficient, out performs many of its variant
Area of Application	Quadratic assignment problem, Travelling salesmen problem, digital image processing	Path generation, test data generation, nano electronic technology	Engineering design and classification	Crowd simulation, controlling nanobots for cancer detection

The research presented in this paper is intended to address the challenge of improving the clustering accuracy and performance to predict the heart disease and liver disease and providing timely response in predicting the disease. Briefly the important research functions are therefore stated as; various datasets are used in the proposed classifier and prediction technique. And a clustering techniques help in developing the prediction model so as to predict accurately the risk of heart disease.

The existing clustering framework requires repeated re-clustering and cluster size specifications, when the data with an incrementally grows. This can be computationally demanding for large uncertain data sets. To address this problem, an effective feature selection and clustering method is proposed. An alternative way to lower the computational cost is to reduce the number of iterations by applying the effective feature selection process, which selects a set of points to as weighted features from large and uncertain dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

III. PROPOSED SYSTEM

Application of data mining in analyzing the medical data is a good method for investigating the existing relationships between variables. Nowadays, data stored in medical databases are growing in an increasingly rapid rate. It has been widely recognized that medical data analysis can lead to an enhancement of health care. The primary objective of the research work is the effective development of prediction model using various clustering techniques to cluster the liver and heart disease. It also shows that data mining can be applied to the medical databases to group or cluster the data with reasonable accuracy. The following are the objectives leading to achievement of the primary objective mentioned:

- To generate a best hybrid meta-heuristic technique this can help in predicting the risk of heart and liver disease with various attributes.
- To recognize and cluster patterns in multivariate patient attributes.
- To predict the class score based on that, the mild and extreme of the disease can be identified.
- To improve the clustering accuracy by utilizing improved hybrid techniques.

To improve the clustering performance, we propose a WPCA (Weighted Principle Component Analysis) [5] based feature selection approach is used and on two types of clinical datasets. It utilizes a fusion based technique for liver and heart disease clustering and to predict the severity of heart disease in patients. The system proposes a new iterative approach, which concentrates on the effective feature selection using PCA (Principle Component Analysis) and effective clustering using improved BAT algorithm. The followings are the main contributions of the proposed work.

- In this paper, we implemented a new WPCA and improved BAT algorithms with the use of effective weighted features. The system introduces a new Liver and Heart disease Clustering algorithm with WPCA technique.
- The improved pre-processing technique calculates the Similarity & Dissimilarity probability Distribution of the cluster by applying a new Combined KL Divergence & Shannon Entropy Distance Measures. This performs the effective pre-processing steps in the given high dimensional medical data sets.
- We also created a new advanced clustering for fast disease analysis. The system developed with the intension of high accuracy and less training overhead.

Fig 1.0 the overall process of the proposed work

Liver and Heart disease clustering and prediction of the class score using a renovation algorithm which is a combination of PCA (Weighted Principle Component Analysis) and improved BAT. In the proposed system, WPCA is used for feature selection and dimensionality reduction and improved BAT algorithm for disease clustering and prediction. The optimized PCA algorithm has been expanded with the new optimal clustering algorithms, which can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time.

A. Data selection:

The real world data is incomplete, noisy and inconsistent. Data pre-processing routines attends to filling the missing values, smooth the noisy data while identifying outliers and correct inconsistency in data. Noise is a random error or variance in a measured variable. Binning can be used for noise removal. Binning methods smooth a sorted data value by consulting its neighborhood. The sorted values are distributed into a number of buckets or bins. At the time of data selection, each value of the attribute is replaced by the mean value of the selected attribute. And this also includes the minimum and maximum values of each attribute values and this will be used as boundaries. Each value is then replaced by the closest value for every attribute.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

B. Feature Selection using Principle Component analysis:

Feature selection is a process commonly used in the second stage of the proposed work, wherein a subset of the features available from the data is selected for application and that will be applied into the learning algorithm. The finest and suitable subset will be selected, that should contain the least number of dimensions. If the dimensions are less the accuracy of the classifier will be high. This discards the remaining, unimportant dimensions using PCA. This is an important stage after pre-processing and is one of two ways of avoiding the curse of dimensionality in the health care domain.

There are two approaches in feature selection namely forward selection and backward selection in the enhanced PCA. Forward Selection starts with no variables and adds them one by one in every iteration; this will reduces the error at the time of clustering. Backward Selection starts with all the variables and removes them one by one. In the feature selection process, all subset selection evaluates the features as a group for suitability. Subset selection algorithms can be segmented into Wrappers, Filters and Embedded. In this research work, WPCA, IBAT Algorithm, have been implemented which are explained in detail in this section.

Algorithm: WPCA+IBAT Algorithm:

Input: patient dataset
Output: optimal feature and Clusters result with risk level and performance result
Steps:

1. Taking the whole dataset –calculate KLSE
2. Find initial component
3. Compute the *d*-dimensional mean vector
4. Compute the covariance matrix of the original or standardized d-dimensional dataset X (here: d=3); alternatively, compute the correlation matrix.
5. Eigen-decomposition: Compute the eigenvectors and eigenvalues of the covariance matrix (or correlation matrix).
6. Sort the eigenvalues in descending order.
7. Choose the k eigenvectors that match to the k^{th} largest Eigen values, here k is the number of features of the new feature subspace ($k \leq d$).
8. Construct the projection matrix W from the k selected eigenvectors.
9. Transform the original dataset X to obtain the k dimensional feature subspace Y ($Y=WT \cdot X$).
10. Return features for clustering

Algorithm1: Feature selection steps

C. Clustering process:

IBAT algorithm is an improved meta-heuristic algorithm that overcomes the optimization problem. After implementing the weighted PXA, the improved bat algorithm, (IBAT) is used. This algorithm is based on BAT, which is developed on echo location behaviour of micro bats. It is based on three important rules. For sensing distance, IBAT uses its‟ echolocation capacity. It also uses echolocation to differentiate between food and prey and background barriers even in the darkness. Bats used to fly randomly with some characteristics like a velocity, fixed frequency and loudness to search for a prey. But in the IBAT, it fly based on the weighted feature generated from the WPCA. It also features the variations in the loudness from a large loudness to average loudness. Bats find the prey using varying wavelength and loudness while their frequency, position and velocity remains fixed. They can adjust their frequencies according to pulse emitted and pulse rate.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Algorithm: Clustering using IBAT

1. Choose initial variables from WPCA result
2. Evaluate the optimalility of each individual in the population using

$$E = \sum_{k=0}^{n} \binom{n}{k} xF^k$$

3. Repeat until termination: (time limit or sufficient result achieved)
   a. Select optimum ranking individuals to reproduce
   b. load next feature based on the feature priority new generation through pulse rate
   c. Evaluate the individual solution
   d. Replace worst ranked part of loudness with offspring
   Choose initial population from the ranked list
4. IBAT sends *m* optimal solutions chosen to the *m attributes.*
5. upon Receiving a Converged Result from the WPCA, the System Stops Execution

Algorithm2: Clustering using IBAT

The proposed system performs the prediction model based on the above IBAT algorithm. The proposed system successfully clusters with optimal feature selection the liver and Heart disease based on the given dataset. The system also predicts the score for the chance of Heart disease based on the boundary calculation. The proposed system implements an optimal clustering which does not depend on the features from WPCA completely. The system performs the statistical properties to evaluate the score of every attribute. The system finally provides the prediction accuracy over the given dataset.

IV. IMPLEMENTATION AND RESULTS

A. DATASET

Two standard clinical datasets of varying sizes and characteristics were obtained from UCI Machine Learning Repository is used in this experiment. The details of the datasets are as follows: The experiment used two datasets for liver. The first standard liver dataset from UCI Machine Learning Repository is used to discriminate healthy people from those with liver disease, according to class attribute which is set to either 0 for healthy and 1 for liver disease. This dataset contains 19 attributes and 1 categorical valued class variable and 106 records. The second data set is used to diagnose the heart disease. The dataset consist of 270 instances collected from all UCI repositories. Using some synthetic dataset a subset is used to evaluate the proposed method. We perform the experiment on the Mayo Clinic patient data obtained during the study period from 1/1999 to 12/2004 with follow-up information available until the summer of 2010. Another dataset used in this study is the Cleveland Clinic Foundation, which is named as Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 13 attributes. The experiment takes Heart disease dataset from UCI repository. The dataset contains 13 attributes considered are: age, sex, FBS (fasting blood sugar > 120 mg/dl), chol (serum cholesterol in mg/dl), restecg (resting electrocardiographic results), trestbps (resting blood pressure), thalach (maximum heart rate achieved), exang

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

(exercise induced angina), slope (the slope of the peak exercise ST segment), oldpeak (ST depression induced by exercise relative to rest). There are a total of 750 patient records in the database. Based on the two real world dataset shown in fig 2.0, 3.0, liver and co morbid conditions associated with liver were assessed.

age	sex	chest_pain	resting_blood_	serum cholestoral in mg/dl	fasting blood sugar > 120 mg/dl	resting electrocardiog results	maximum heart rate achieved	exercise induced angina
70	1	4	130	322	0	2	109	0
67	0	3	115	564	0	2	160	0
57	1	2	124	261	0	0	141	0
64	1	4	128	263	0	0	105	1
74	0	2	120	269	0	2	121	1
65	1	4	120	177	0	0	140	0
56	1	3	130	256	1	2	142	1
59	1	4	110	239	0	2	142	1
60	1	4	140	293	0	2	170	0
63	0	4	150	407	0	2	154	0

Fig 2.0 Heart dataset

In this paper, Principal Component Analysis is used for Feature extraction and IBAT Algorithm, a hybrid method is used for clustering. Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. The huge size in data may affect the algorithm performance. So, Feature extraction is a mandatory process for constructing combinations of the variables to get around these problems. Best results are achieved when the features are constructed from the effective feature selection process and a set of application dependent features. Feature extraction is implemented using the Principal Component Analysis method and Linear Discriminate Analysis. Weighted PCA and IBAT Algorithm has been successfully applied for various problems.

DATA SIZE	PCA	WPCA
100	0.4	0.098
200	0.3	0.105
300	0.45	0.13
450	0.48	0.12
500	0.5	0.3

Table 1.0 False Positive Rate comparison table

From the results shown in Table 1.0, it is found that WPCA Based Clustering Optimization increase the True Positive Rate significantly and reduce the False Positive Rate to an acceptable extent.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

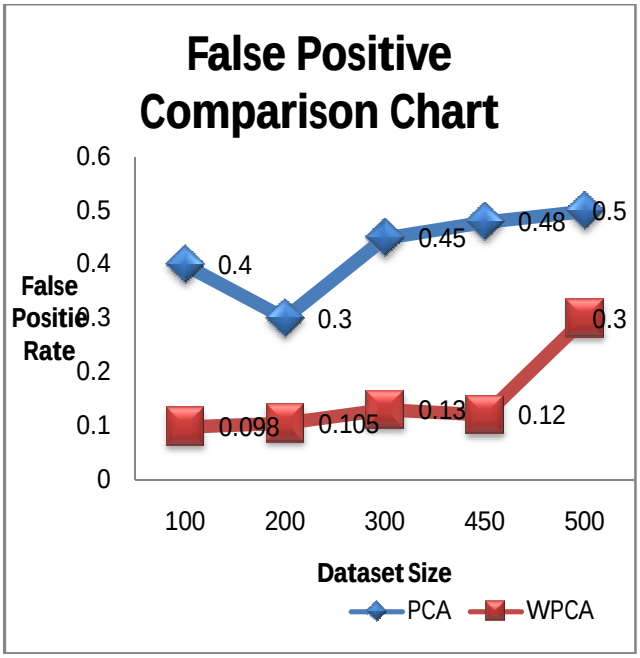

Figure 3.0 Performance Measure in terms of False Positive Rate

As in fig 3.0 The false positive rate of the proposed system is quite high, because some normal classes in the additional data merged could be clustered as a disease, but only the weighted features are used in grouping. The reduction in false positive rate of the proposed system is mainly due to the WPCA and IBAT process.

Precision: Precision for a class is the number of true positives (i.e. the number of instances correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class). The equation is:

$$\text{Precision} = TP / (TP + FP)$$

Recall: Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been.) The Recall can be calculated as:

$$\text{Recall} = TP / (TP + FN)$$

Accuracy: The percentages of the predicted values are match with the expected value for the given data. The best system is that having the high Accuracy, High Precision and High Recall value. The performance of the proposed system is tested with the 270 instances, from each instance the precision and recall values are gathered and that is plotted in the fig 2.0. With help of the confusion matrix values measurement of the precision and recall values are calculated and plotted as a graph below: It is observed that the performance is very promising compared to the existing methods that have been explored in the previous chapter. The next chapter deals with the presentation of the conclusion and enhancements.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Table: 2.0 Performance comparison table

Type	BAT	WPCA_IBAT
Feature selection Time(s)	8	3.4
Clustering Time (s)	5	2.6

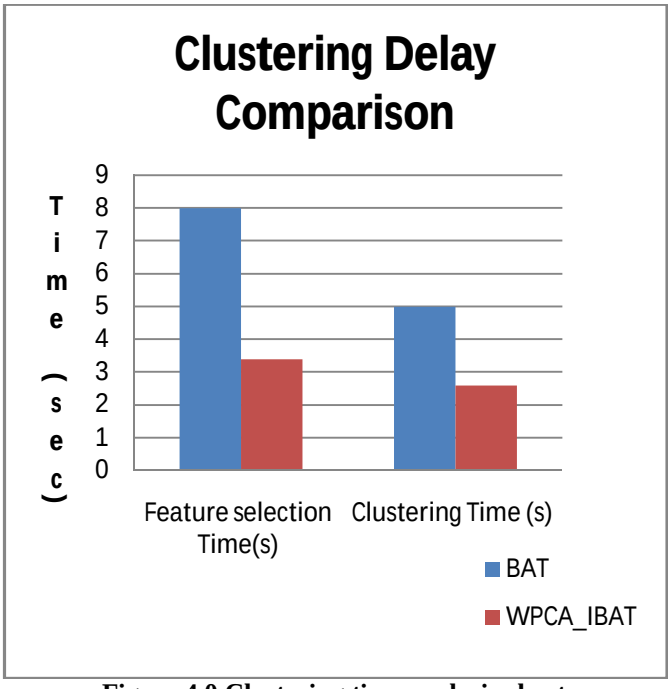

Figure 4.0 Clustering time analysis chart

As in fig 4.0 the clustering delay of the proposed system is quite reduced than the existing algorithms. Due to the dimensionality reduction using WPCA and IBAT reduces clustering delay.

VI. CONCLUSION

In this paper, we proposed a new clustering and prediction scheme for liver and Heart disease data. The system studied the main two problems in the literature, which are clustering accuracy and clustering delay. The study overcomes the above two problem by applying the effective enhanced weighted component with IBAT algorithm. The PCA represents with the effective splitting criteria which has been verified by the IBAT algorithm. The system effectively identifies the disease and its sub types, the sub type which is referred as the percentage of class.

The experimental results are evaluated using the two set of datasets. The experimental result shows that integrated extended weighted component with IBAT algorithm shows better quality assessment compared to traditional PCA and WFF techniques. From the experimental results, the execution time calculated for clustering object is almost reduced than the existing system. The proposed framework model can be used to analyse the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

research in the field of data mining in other dataset and use of other clustering algorithms. As further work, use this model as a functional base to develop an appropriate data mining system for clustering performance.

REFERENCES

[1]. I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.
[2]. G. D. Magoulas and A. Prentza, "Machine learning in medical applications," *Mach. Learning Appl. (Lecture Notes Comput. Sci.)*, Berlin/Heidelberg, Germany: Springer, vol. 2049, pp. 300–307, 2001.
[3]. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.
[4]. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of clustering techniques." (2007): 3-24.
[5]. Wang, Xuechuan, and Kuldip K. Paliwal. "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition." *Pattern recognition* 36.10 (2003): 2429-2439.
[6]. Ruben, D.C.J., *Data Mining in Healthcare: Current Applications and Issues.* 2009.
[7]. Porter, T. and B. Green, *Identifying Liver Patients: A Data Mining Approach.* Americas Conference on Information Systems, 2009.
[8]. Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, *Data mining techniques for cancer detection using serum proteomic profiling.* Artificial Intelligence in Medicine, Elsevier, 2004.
[9]. Das, R., I. Turkoglu, and A. Sengur, *Effective diagnosis of heart disease through neural networks ensembles.* Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.
[10]. Yang, X.-S.: "Nature-Inspired Metaheuristic Algorithms ". Luniver Press, (2008)
[11]. X.S. Yang, "Firefly algorithm, stochastic test functions and design optimization," International Journal of Bio-Inspired computation, vol. 2, no. 2, pp. 78-84, 2010.
[12]. X.S. Yang, "Firefly algorithm, levy flights and global optimization," in Research and Development in intelligent systems XXVI. Springer, 2010,pp.209-218.
[13]. Johnson, Eric G., et al. "Advantages of genetic algorithm optimization methods in diffractive optic design." *Critical Review Collection*. International Society for Optics and Photonics, 2017.
[14]. X.-S. Yang, S. Deb, "Cuckoo search via L´evy flights", in: Proc. Of World Congress on Nature & Biologically Inspired Computing (NaBIC2009), December2009, India. IEEE Publications, USA, pp. 210-214(2009).
[15]. P. Musikapun, P. Pongcharoen, " Solving Multi-Stage Multi- Machine Multi-product Scheduling Problem Using Bat Algorithm", 2nd international Conference on Management and Artificial Intelligence, IPEDR Vol.35, 2012