JOURNAL OF MECHANICS OF CONTINUA AND
MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-7, July (2020) pp 282-293 ISSN (Print) 0973-8975

A SURVEY ON VARIOUS CLUSTERING ALGORITHMS USING NATURE INSPIRED ALGORITHMS

Mohammed Ali Shaik[1], **P. Praveen**[2]

[1]Assistant Professor, Member Center for Embedded & IOT systems, S R Engineering College, Warangal, Telangana State, India

[2]Associate Professor, S R Engineering College, Warangal, Telangana State, India

[1]niharali@gmail.com, [2]prawin1731@gmail.com

Corresponding Author: Mohammed Ali Shaik

https://doi.org/10.26782/jmcms.2020.07.00023

Abstract

K-means clustering algorithm and its variants have many drawbacks and one of the major one is getting stuck at local optima while calculating centroids over random values. Algorithms that optimize computation are iterative in nature for speeding up the process of creation or search of data by multiple search agents. Swarm intelligence (SI), is a primary aspect of artificial intelligence that comprises of high complexity problems and proposed solutions that are sub-optimal and achievable in a given time span. SI adopts cooperative character of an organized group of animals that are formed on the phrase: strive to survive and in this paper we provide a detailed survey of eight different SI algorithms that are related to insect and animal based algorithms and provides initial understanding and exploring of technical aspects of algorithms.

Keywords : Swarm intelligence; Machine learning, K-means, Bio-inspired algorithms, Intelligent algorithms, Literature review, Nature-inspired computing.

I. Introduction

"Swarm intelligence (SI)" [IV] or "bio-inspired computation" is a subset of artificial intelligence (AI) [XII] which is identified to be an emerging field by most of the current day researchers, it is incepted by Gerardo Beni and Jing Wang in 1989 [VI] as a part of cellular robotic system development process. One of the major reasons for popularity of SI-based algorithms is flexibility and versatility in nature and some of the other features are individual learning potentiality and implement ability towards external variations. These key features attracted massive interest that lead to adoption of several applications in these areas.

"Swarm intelligence" has gained popularity in recent times with drastic increase in prominence of NP-hard problems [XII] that evaluates global optima which is almost impossible to implement in present day scenarios as the number of possible solutions that merely exists in such a problems space that comprises of or consists of a broader scope. In most of the circumstances getting a implementable and acceptable solution in a given time span limitation which is a important aspect [XXV]. Where SI provides us with better alternatives in resolving nonlinear design issues [XIII] with real time applications [IV] “in almost all areas of sciences or engineering and industries or from data mining to optimization or computational intelligence or business planning or in bioinformatics and in industrial applications” [XII, VI, VII, XXII].

A considerable collection of “nature-inspired optimization” methods are meta-heuristics [I] that have been emerged very recently that comprises of designs mimicking called as “swarm” behavior that exhibits scenarios of living creatures [XVI]. Each of the searching agent specify a precise combination of centroid position evaluated by performing search operation for attaining optimality in their own ways as they communicate with one another in a guided form towards attaining a global optimization objective.

II. Research Methodology

The proposed research is being comprehensively conducted and is still going as eight of the algorithms have been studied up to a maximum extent where authors have made sincere efforts to enlighten un familiar algorithms and identified reasons to be in their immature development stage that comprises of “genetic algorithms or neural network” [XXII] that have been extensively studied and leads to or into several publications that are made available that supports the same literature and the process of development is leaved in the areas of such less familiar algorithms that have been identified using the major source of “Internet and publications” more specifically “meta-heuristics and nature-inspired algorithms” [IX, XII, XVII, XIX].

Authors have collected information such as the recent development in the algorithms and the later stages comprises of comprehensive literature survey which focuses description and parameters of all algorithm that are chosen in this paper along with better understanding of knowledge. It also provides substantial synopsis of the probable scope of various applications over these algorithms that tend to adopt detailed methodology discussed as above on the eight swarm based algorithms were first of all identified and then broadly classified as “Insect based and Animal based” algorithms are shown in Fig 1

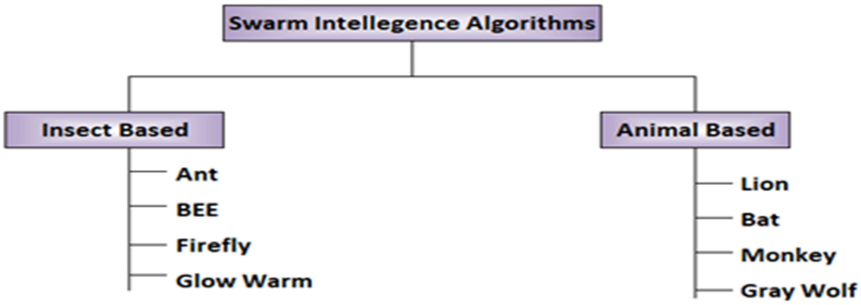

Fig. 1: SI based algorithms used in the present study

"Insect based (IB) algorithms such as bee-inspired and ant colony are the individual IB algorithms and fire fly and glow-worm [II–IV] and individual animal based algorithms such as monkey, wolf, lions and bats" [VI–XI] has been explained for their distinct applications in various domain areas that specify the detailed analysis that has been done by us. In the below sub sequent sections all the eight algorithms of SI are being highlighted with brief enlightening description along with the parameters that are used before initiating the concept of application implementation with distinct modes in various problem domains. Fig1 represents fundamental hierarchy and classification of SI based algorithms that are presented in detail in this study.

III. Enhancing K-Means Algorithms By Insect Based Swarm Intelligence Algorithms :

K-means clustering algorithm results to be obtaining the correct combination of attaining centroid positions. The generated centroids are based on the yield value which *is t*he maximum quality called as global optimum where the characteristics or properties that comprises of maximum or minimum similarities in clustering.

One of the drawbacks of K-means clustering is that the clusters do not always cover global optimum as that as partition-based clustering algorithm are, where the starting partition is made using random centroid values without guaranteeing subsequent convergence with best possible clustering result obtained. All the final clusters consists of local optima leads to prevention of further exploration to attain best results [XII] which requires many rounds to be executed with possible distinct random initiations with generated centroid values.

Almost all the drawbacks of K-means algorithm can be handled effectively with the nature-inspired optimization algorithms, which are computed stochastically from start to its convergence based on the algorithm implemented for searching the agents based on centroids principal that relocate iterations with inclined values that enable the new clusters to be generated with well formed results to achieve the optimal configuration of centroids using:

$$W_{i,j}=\{0,\ x_i \notin \text{cluster}_j \mid 1,\ x_i \in \text{cluster}_j\} \qquad (1)$$

Where $W_{i,j}$ is the membership of data point x_{ij}, and i,j denote clusters.

$$\boldsymbol{centroid_{i,j}} = \frac{\sum_{i=1}^{s} w_{i,j}\, x_{i,v}}{\sum_{i=1}^{s} w_{i,j}}, \boldsymbol{j = 1, \ldots \ldots K, v = 1, \ldots, k} \otimes \mathrm{D} \qquad (2)$$

Where S represents number of search agents in the whole population, K is maximum number of clusters and j is the current cluster under execution over a data set size D.

Ant Colony Optimization (ACO) Algorithm [5]

As per the real-life behavior of "ant colonies" [II] are inspired to form of "Artificial ant colony systems or algorithms" where every single ant agent is identified as to be biological entity [XIII] whose capabilities are predefined and fixed by considering an ant colony into account [XIV] as such artificially intelligent algorithms [XXI] possess inherent behavior that are required to implement the system.

Most of the common applications in the present day include the development process of intelligent solutions [XV] that performs simplified process towards transportation of heavy goods and materials [XVI] by identifying the possible shortest path between the destination and source agents [XVII]. The process of performing communication between various groups of ants that is indirect and self-organized as most of the exclusive ants will not perform direct communication with the other ants or members of the pack, this process of communication in ant colonies is performed by using a "chemical substance known as pheromone" [XVI-XX] where each ant agent spits while moving or traversing in a path and all other ant agents simply follow the "pheromone" substance and further reinforce it by adding their own pheromone.

The group of ants communicate by "stochastic local decision policy" application that depends over basic factors such as "pheromone and heuristic values as the intensity of pheromones" spitted will be equal to total number of ants walking one after another in a particular trail or path with the attained heuristic value that is generated as it is problem based [XVIII]. The process includes several iterative steps executed in a serial order from a sample of solutions where the heuristic values are updated as and when an ant obtains optimal solution. After generating multiple paths or after distinct iterations the ant agent selects to track and follow the reinforced pheromone trail and rejects the rest of the paths that has lesser pheromone intensity when compared and in this process comprises of many iterations that generates acceptable results in a feasible amount of searching time span and this algorithm are thoroughly illustrated in Table 1.

Table 1: Parameters used in Ant colony optimization algorithm:

Parameter	Description
N	Population of Ants in an Ant colony
M	Number of Ant Colonies
C	Cluster
K	Number of groups of clusters used
S	Number of samples
$W_{s,k}$	$W_{S,K}$={0, $x_S \notin$ cluster $_K$ | 1, $x_S \in$ cluster $_K$ }, where s $\in$ S and k $\in$ K
$x_{i,j}$	The adoptable solution in matrix with indices i, j, where i belongs to N and j belongs to K such that K $\otimes$ D
J	Cost function: $\boldsymbol{J = \sum_{i=1}^{n} min_{j=1}^{k} \|\|x_i\ c_j\|\|\ \ ^2}$
centroid (k, d)	The centroid function in matrix [k,d], where k belongs to K and d belongs K $\otimes$ D

Bee Inspired Optimization Algorithm

The artificial bee colony algorithm is also called as one of the versatile meta-heuristic method that employees a technique by an intelligent swarm that is used by bees to uniquely identify their locations of food by using their communication method that comprises of selecting nest location, task allocation or implementation, performing reproduction, doing dance, etc, are used as inputs to update the algorithm as per problem requirements in an iterative manner. The artificial bee colony algorithm searches data using best-fit solution over the large number of data while resolving critical problems. These bees are classified into three categories: employed, onlooker and scout bees. Where every type of bee performs their own task where [XIX]

Scout bees will perform random search for identification of new food sources and marks it with fitness quotient and then if fresh food source is represented by any of the employed bee with major degree of fitness and the fresh source is modified to the new location by ignoring the scout identified food source, as the main task of employed bee is to continuously update the food source database [XX]. The task of onlooker bees is to identify and evaluate fittest based on quantity of presence of food or with any other parameter such as distance, etc, if the bees cannot improve fitness quotient of food source then all the solutions provided by employed bee or onlooker bees is nullified and the parameters used in this algorithm are illustrated in Table 2.

Table 2: Parameters used in Bee colony optimization algorithm:

Parameter	Description
X	number of Bees
I	Food sources
D	Variables present in optimization problem
K	Number of groups of clusters used
S	Number of samples
x_{minj}	Lower restriction for food source fitness
x_{maxj}	Higher restriction for food source fitness
$x_{i,j}$	Represents each food source: $=x_{minj}+rand[0,1](x_{minj}+x_{maxj})$
$V_{i,j}$	Represents position updation by: $v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj})$
centroid(k, d)	Te centroid function in matrix [k,d], where k belongs to k, d belongs $K \otimes D$

Firefly Based Algorithm

Firefly algorithm is very good at handling complicated problems that consists of equality or inequality based criteria where algorithm treats every firefly as multi-modal function with higher efficiency when compared with the "ant and bee swarm algorithms". This algorithm also adopts basic random population based search over a group of distinct solutions that results with maximum convergence and error-free outcome [XXI]. Fireflies communicate with flashing or glowing signals about identified prey or identification of mates or to perform communication as similar to other swarm intelligence characteristics using their own self organizing & decentralized decision making capability as the glowing capability of firefly represents its fitness level, though in conventional firefly algorithm attract one another in similar as they are considered to be unisex which also represents potential "candidate solution" and the parameters used in this algorithm are illustrated in Table 3.

Table 3: Parameters used in firefly optimization algorithm:

Parameter	Description
N	Total population of fireflies
M	The total number of distinct time stamps
K	The number of distinct groups of clusters used
S	The number of samples
D	Total number of distinct attributes in a dataset
W(s,k)	$W_{S,K}=\{0, x_S \notin \text{cluster}_K \mid 1, x_S \in \text{cluster}_K \}$, in which s ∈ S and k ∈ K
x(i, j)	Best fit solution generated using matrix [i, j], in which I ∈ N and j ∈ K ⊗ D
centroid(a, b)	Calculation of centroid in matrix [a,b], where a belongs to K and b belongs K ⊗ D
Cal_mat(a,b)	Calculation of classification matrix [a,b] where a ∈ N and b ∈ S

Glow Worm Based Algorithm

"Glow-worm-based algorithm is one of the swarm intelligence-based algorithm" that optimizes multi modal functions that are based on behavior of glow-worms using machine learning systems. Glow-worms possess a chemical called "luciferin emission" that allows a worm to glow at various intensities [IV, V] using which they communicate with one another and the luciferin that is induced will attract mates in the process of reproduction or for feeding the preys. If the intensity of luciferin is more glow worm attracts more prays or glow worms because more brighter the intensity of glow represents more the value of fitness of the worm. Initially glow warms are randomly identified and selected to create an artificial swarm environment. Each and every worm agent represent optimization problem by utilizing search domain for identifying the target neighbors and moves into a specific direction based on the intensity of luciferin picked up from other glow worm if the intensity is more than itself. In the algorithm each glow worm individually considered to be swarm as it is assigned with an objective function [XXIII] and luciferin [XXII] intensity level based on its present location this is obtained and the parameters used in this algorithms that are illustrated in Table 4.

Table 4: Parameters used in glow worm optimization algorithm:

Parameter	Description
N	total population of glow worms
M	Total number of time stamps
K	Total number of groups of clusters used
S	Total number of samples
D	Total number of attributes in a dataset
L	The Luciferin level
$Lj\ (t-1)$	The previous luciferin level
P	The luciferin decay constant ($\rho \in (0, 1)$);
Γ	The luciferin enhancement fraction
$F(pj\ (t))$	objective function value for glowworm j at current glowworm position (pj);
T	Current iteration
$Lj\ (t)$	Luciferin level= $(1-\rho)\ Lj\ (t-1)+\gamma F(pj\ (t))$.
x(a, b)	The best solution in matrix [a, b] where $a \in N$ and $b \in K \otimes D$
centroid(a, b)	Centroid function in matrix [a,b] in which a belongs to K and b belongs $K \otimes D$

IV. Comparative Study of Optimization Algorithms

Table 5: Techniques adopted by various authors over the optimization techniques

Optimization algorithm	Authors	Technique	Description
Ant Colony	Abraham A, et al. [V]	“Artificial Ant Colonies”	“Classification techniques and rules are applied in Artificial ant colonies“
	Maniezzo V [VI]	“Quadratic Assignment problem”	“Solved combinatorial optimization problems”
	Forsyth P, et al. [VII]	“Bus driver scheduling”	“Provide solution for best scheduling of drivers such that drivers are fully utilized”

	Davidovic, T., Jak, et.al.[VIII]	“General BCO (GBCO)”	“Proposed use of the asynchronous execution strategy in two ways: centrally coordinated knowledge exchange and second is non-centralized parallelism”
	Plamenka B., Veska G. [IX]	“For the case study of the influenza virus sequences”	“Hybrid parallel implementation utilizing MPI and Open-MP provides considerably better performance than the original code + It allows researchers to perform simulations with very large amounts of data in the field of bioinformatics to conduct their experiments on even more powerful supercomputers”
Fire fly	Mishra A, [10]	“Grey scale image watermarking”	“Proposed a noval optimized gray scale based image watermarking system using DWT-SVD and firefly algorithms”
	Rahmani A, et al. [XI]	“Capacitated facility location problem”	“Proposed a hybrid firefly based genetic algorithm for the capacitated capability location problem”
	Verma OP, et.al. [XII]	“Heart disease prediction”	“Opposition and dimensional based modified firefly algorithm for heart disease prediction”
	Apostolopoulos, et.al. [XIII]	“Load dispatch problems”	“Proposed an noval application using firefly algorithm for resolving the economic emissions using load dispatch problem”
Glow-worm	Krishnanand KN, et al. [XIV]	“Parameter optimization multi-modal search”	“Proposed a new method for optimizing multi-modal functions”
	Krishnanand KN, et al. [XV]	“Searching of multiple local optima for multi-modal functions”	“Proposed a method that simultaneously capture data based on multiple local optima over a multimodal function implementation”
	Krishnanand KN, et al. [XIV]	“Signal source localization”	“Proposed a glowworm based swarm optimization using enhanced multi robot system for performing signal source localization process”
	Senthilnath J, et al. [XVII]	“Image processing”	“Proposed a hierarchical clustering algorithm for land cover mapping using satellite images”
	Zhou YQ, et al. [XVIII]	“Travelling salesman problem”	“Proposed a separate glowworm bsed swarm optimization algorithm for resolving a TSP problem”

Based on the above table where we compared all the eight optimization algorithms where each algorithm has been considered with seven best papers from various authors and our analysis is:

- "Ant colony based algorithm" has vast variety of applications that are based on ant colony-based algorithm and potential applications have been identified in data mining to implement classification and it has a difficult theoretical analysis to be don, ants perform dependent sequences of decisions, uncertain time to search and more theoretical and research has to be done.
- "Bee colony based algorithm" finds applications to be single objective numerical value optimization technique. The major disadvantage in this technique is it the search operation performed is very limited initially as it adopts normal distribution where each sample is to be initialized in every iteration.
- "Fire fly algorithm" comprises of host applications that are based on multi model optimization process and can be used efficiently in NP based hard problems but the disadvantages in this technique is it is hard to code as it is multi model and comprises of very few literature examples.
- "Glow worm algorithm" are trivial in terms of exploration in theory and implementation and the disadvantages include dissimilarity in objective function as lucriferin value has to be enhanced consistently.
- Gray wolf algorithm has many advantages such as multi layer perception and to provide training on "q-Gaussian radial basis over functional-link nets" and has the capability to solve complex problems, few of them includes:
- Dispatch problems that are economically feasible
- Dispatch problems that are based on combined economic emission
- Best possible allocation process of STATCOM devices using the power system grid
- For solving evolutionary population dynamics
- Multiple input and multiple output contingency based management problems

V. Future Research Directions

"Insect-based algorithms such as ant colony and bee colony then firefly-based algorithm and bat-based algorithm are from the animal category" are the former group of algorithms that are extensively used to optimize data. Immense research has been carried out in a manageable form where the content is explored which supports with threshold number of feasible references. It needs huge research to be done in other algorithms such as "lion based or wolf" based algorithms because these algorithms are novel. Our research focus on all the above said algorithms but more on the novel algorithms as they are considered to be more advanced and simplified as these techniques are time effective over complicated computation problems. Further our research has to explore more on individual optimization algorithms and their sub domains for performing review in terms of application and their scope.

VI. Conclusion

The research in this paper will present a detailed review and parameters used with formulas on selective well known swarm intelligence and novel algorithms that are used to optimize the searching capability either for pray of for some other reasons as specified . In this paper we have taken into consideration only “insect based and animal based algorithms” that is popular among the current day researchers in the area, for which we have studied and explored more than 90 papers in the area.

References

I. B. Ozden, S. Ramaswamy, A. Silberschatz. Cyclic associ-ation rules. Proceedings of the 15 th International Conference on Data Engineering. 1998, 412-421.

II. D. R. Li, S. L. Wang, W. Z. Shi et al. On Spatial Data Mining and Knowledge Discovery [J]. Geomatics and Information Science of Wuhan University, 2001, 26(6): 491-499.

III. Dr. Seena Naik, “An Effective use of Data Mining Techniques to Creation”, International Journal of Advancement in Engineering, OCT, 2016, volume: 3, Edition: 10, pp.157-163, ISSN: 2349-3224.

IV. D. Ramesh, Syed Nawaz Pasha, G. Roopa, "A Comparative Analysis of Classification Algorithms on Weather Dataset Using Data Mining Tool", Oriental Journal of Computer Science and Technology, DEC, 2017, Volume:10, Issue:4, Pp.1-5, ISSN:0974-6471.

V. Forsyth P, Wren A (1997) an ant system for bus driver scheduling. Research Report 97.25, University of Leeds School of Computer Studies

VI. J. Han, G. Dong, Y. Yin. Efficient mining of partial periodic patterns in time series database. In Proc. 1999 Int. Conf. Data Engineering (ICDE'99), pages 106-115, Sydney, Australia, April 1999.

VII. Ji, Xue, et al. "PRACTISE: Robust prediction of data center time series." International Conference on Network & Service Management 2015. G. Box, G. M. Jenkins. Time series analysis: Forecasting and control, Holden Day Inc., 1976.

VIII. Kuo RJ, Chiu CY, Lin YJ (2004) Integration of fuzzy theory and ant algorithm for vehicle routing problem with time window. In: IEEE annual meeting of the fuzzy information, 2004. Processing NAFIPS’04, vol 2, pp 925–930. IEEE

IX. Mishra A, Agarwal C, Sharma A, Bedi P (2014) Optimized gray-scale image watermarking using DWT-SVD and firefly algorithm. expert syst appl 41(17):7858–7867

X. Mohammed Ali Shaik, “A Survey on Text Classification methods through Machine Learning Methods”, International Journal of Control and Automation, Vol. 12, No.6, (2019), pp. 390 – 396.

XI. Mohammed Ali Shiak, “A Survey of Multi-Agent Management Systems for Time Series Data Prediction”, International Journal of Grid and Distributed Computing, Vol. 12, No. 3, (2019), pp. 166-171.

XII. Mohammed Ali Shaik, “Time Series Forecasting using Vector quantization”, International Journal of Advanced Science and Technology, Vol. 29, No. 4, (2020), pp. 169-175.

XIII. Mohammed Ali Shaik, S Narsimha Rao, Abdul Rahim, “A SURVEY OF TIME SERIES DATA PREDICTION ON SHOPPING MALL”, Indian Journal of Computer Science and Engineering (IJCSE),Vol. 4 No.2 Apr-May 2013, ISSN : 0976-5166, pp. 174-184.

XIV. Mohammed Ali Shaik, P. Praveen, Dr. R. Vijaya Prakash, "Novel Classification Scheme for Multi Agents", Asian Journal of Computer Science and Technology, ISSN: 2249-0701 Vol.8 No.S3, 2019, pp. 54-58.

XV. Nakamura RY, Pereira LA, Costa KA, Rodrigues D, Papa JP, Yang XS (2012) BBA: a binary bat algorithm for feature selection. In 2012 25th SIBGRAPI conference on graphics, patterns and images. IEEE, pp 291–297

XVI. P. Praveen, C. J. Babu and B. Rama, "Big data environment for geospatial data analysis," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2016, pp.1-6.doi: 10.1109/CESYS.2016.7889816.

XVII. Praveen P., Rama B. (2018) A Novel Approach to Improve the Performance of Divisive Clustering- BST. In: Satapathy S., Bhateja V., Raju K., Janakiramaiah B. (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, vol 542. Springer, Singapore.

XVIII. R. Ravi Kumar, M. Babu Reddy and P. Praveen, "A review of feature subset selection on unsupervised learning," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 163-167.doi: 10.1109/AEEICB.2017.7972404.

XIX. Schoonderwoerd R, Holland OE, Bruten JL, Rothkrantz LJM (1996) Ant-based loadbalancing in telecommunications networks. Adapt Behav 2:169–207

XX. Shekhar, P. Zhang, Y. Huang et al. Trends in Spatial Data Mining. In: H. Kargupta, A. Joshi(Eds.), Data Mining: Next Generation Challenges and Future Directions[C]. AAAI/MIT Press, 2003, 357-380.

XXI. Socha K, Knowles J, Sampels M (2002) AMAX-MIN ant system for the university timetabling problem. In: Dorigo M, Di Caro G, Sampels M (eds) Proceedings of ANTS2002—third international workshop on ant algorithms. Lecture notes in computer science, vol 2463. Springer, Berlin, Germany, pp 1–13

XXII. T. Sampath Kumar, B. Manjula, D. Srinivas, "A New Technique to Secure Data Over Cloud", Jour of Adv Research in Dynamical & Control Systems, 11-Special Issue, July 2017.

XXIII. T. Sampath Kumar, B. Manjula, Mohammed Ali Shaik, Dr. P. Praveen, "A Comprehensive Study on Single Sign on Technique", International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238E-ISSN:2207-6360, Vol-127-June-2019