Journal of Theoretical and Applied Information Technology
15th April 2017. Vol.95. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NOVEL ALGORITHM FOR BIG DATA CLASSIFICATION BASED ON LION OPTIMIZATION

[1]NAVNEET, [2]NASIB SINGH GILL

[1]Research Scholar, Department of Computer Sc. and Applications, MDU, Rohtak, Haryana, INDIA
[2]Professor and Head, Department of Computer Sc. and Applications, MDU, Rohtak, Haryana, INDIA
E-mail: [1]navneet.khatri@gmail.com, [2]nasibsgill@gmail.com

ABSTRACT

This paper develops a novel big data classification algorithm based on a nature inspired meta-heuristic algorithm (lion optimization algorithm). Lion optimization algorithm is an optimization algorithm based on the hunting and social behaviour of the lion. The developed algorithm uses the K-mean clustering to generate the pride and nomad. Then the hunting and migration behaviour of the lion is repeated to change pride and optimize the process. The proposed calculation is dissected by adding the proposed calculation to the WEKA library on the Intel i5 @ 2.67 GHz utilizing the Eclipse IDE. The calculation is examined on the datasets having 400 occasions with 25 qualities and 32561 examples with 15 properties. The algorithm has been analyzed on different datasets using Tp rate, Fp rate, accuracy, recall and f-measure as parameters. The result analysis shows the optimization of the algorithm.

Keywords: *Big Data, Lion Optimization Algorithm, Classification, Accuracy, Meta-Heuristic, Nature Inspired Algorithm*

1. INTRODUCTION

During past decade, as dataset grows complexity also increases. This enhanced complexity and large dataset size makes enhance the difficulty to process the data. The main reason for such a large data is the recording of all attribute without knowing their significance in any event. The processing of data firstly needs the extraction of the significant attributes from the available attributes. That's why, the present techniques of data mining of big data firstly performs the dimension reduction step to reduce the number of attributes. Then the reduced dimension datasets is used for the processing. There are several searching techniques to find the significant features from the all features like classical, heuristic and meta-heuristic technique etc. The meta-heuristic techniques have proven their significance to solve such tasks. There are many meta-heuristic algorithms to solve many problems related to scheduling problems, data clustering, image and video processing, tuning of neural networks and pattern recognition etc.

Data mining is process of interpreting data from different context. There are several data mining techniques such as classification, association, clustering, prediction and sequential pattern. Decision tree is a widely used data mining method. In decision theory, a decision tree is a graph of decisions and their possible results, represented in form of branches and nodes. Decision trees is one of the traditional technique used for both classification and estimation tasks and can be used to forecast the conclusion for new samples. [1]

Existing approaches of decision tree are ID3 [2] ,C4.5[3] ,C5.0[4], CHAID, CART[5] are based on different node splitting criteria The CART (classification and regression trees) based on Gini coefficient can only make a binary tree and backward pruning. C4.5 is improved form of ID3 and also based on entropy. The meta-heuristic techniques can be used in solving various classical problems effectively. Classification is the one of such problem which can be solved more effectively by using meta-heuristic algorithms.

The contribution of the paper is to design an optimized classification technique which selects

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the splitting criteria in such a way that impurity is minimized and the accuracy is improved. This is achieved by using a meta-heuristic technique known as lion optimization algorithm, which depicts the attacking and living behaviour of the lion. .

2. LITERATURE REVIEW

Nature is a good guide to solve problem and find most appropriate solution. A number of nature inspired algorithms are meta-heuristic in nature like Genetic Algorithm (GA) [6],Particle Swarm Optimization (PSO) [7],Ant Colony optimization (ACO) [8],Gravitational Search Algorithm (GSA) [9],Bat Algorithm (BA) [10], Lion Optimization Algorithm (LOA) [11], Water Wave Optimization (WWO) [12] etc. The genetic algorithm uses hereditary administrators like crossover and mutation that are connected on guardian trees to create offspring. This process proceeds till a predefined halting criteria relating to ideal group parcel get fulfilled. Particle swarm optimization is a heuristic worldwide improvement technique furthermore a streamlining calculation, which depends on swarm insight. It originates from the exploration on the fledgling and fish run development conduct. The calculation is generally utilized and quickly created for its simple usage and couple of particles required to be tuned. In the fundamental of PSO calculation, particle swarm comprises of "n" particles and the position of every particle remains for the potential arrangement in D-dimensional space. The particles change its condition as indicated by the accompanying three standards: (1) to keep its inactivity (2) to change the condition as indicated by its most hopeful particle position (3) to change the condition as indicated by the swarm's most confident position. GSA depends on the law of gravity and mass cooperation. In the GSA calculation, the searcher specialists are an accumulation of masses which cooperate with each other taking into account the Newtonian gravity and the laws of movement. This calculation depends on the Newtonian gravity: Every particle in the universe pulls in each other particle with a power that is specifically relative to the result of their masses and contrarily corresponding to the square of the separation between them. In both GSA and PSO the streamlining is acquired by specialist's development in the pursuit space, however the development methodology is distinctive. Some imperative contrasts are as per the following: In PSO the heading of an operator is ascertained utilizing just two best positions, pbest and gbest. In GSA, the specialist heading is computed in view of the general power acquired by every single other operator. In PSO, upgrading is performed without considering the nature of the arrangements and the wellness qualities are not imperative in the overhauling technique while in GSA the power is corresponding to the wellness worth thus the operators see the hunt space around themselves in the impact of power.

Xiaofen Lu et. al. designs a classification-assisted differential evolution which reduces the computational cost of the differential algorithm. It basically incorporates the classification technique in the differential evolution algorithm to make work efficiently on classification problems[13]. The existing techniques perform the classification with higher accuracy as compared to traditional algorithms like J48. But sometime these techniques get stuck in local optima. LOA is the one of the latest nature inspired meta-heuristic algorithm discussed in next section. The LOA moves toward the global optima more effectively as compared to existing techniques like ACO, PSO etc [11]. So, this paper proposes a LOA based algorithm for the big data classification presented in the third section of the paper. Then the paper discusses the implementation and result analysis of the algorithm over different datasets. Finally the paper concludes the significance of the work.

3. LOA

Lion optimization algorithm given by M. Yazdani and F. Jolai [11] is a population based meta-heuristic algorithm. This algorithm is inspired from the hunting and social behaviour of the Lion. Lions are categorized in two types of social organizations one is pride and other is nomads. The prides are residents in groups while the nomads move sporadically either in pair or singularly. In the LOA each solution is marked as lion given as $\text{lion} = [s_1, s_2, s_3, \ldots\ldots s_n]$.The cost function C is used to evaluate the cost of each lion. Initially SN solutions are generated randomly in search space n which %P is nomads and rest are divided into c clusters of prides. The %f in each pride is females and remaining are the males. The tertiary of the pride is defined by best visited position of each lion [11].

The hunting style of the pride has been specific to encircle the quarry. A few female of the pride attack the quarry while other move to area of province. The hunting members of pride divide themselves into three groups i.e. center, left and right section. The group uses the opposition based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

learning [14] to attack the quarry. The group with maximum fitness value is considered as the center group and position of the quarry is related to the position of falconers is shown in (1):

$$P_{quarry} = \sum p_{falconers}(s_1, s_2, s_3, \dots\dots\dots s_{nv}) / \text{number of falconers} \quad (1)$$

Here, P_{quarry} and $p_{falconers}$ are the position of quarry and falconer respectively. During the hunting process, the falconers are elected randomly and position updation of the quarry and falconers (left, right and center section) is given by (2), (3) and (4) respectively.

$$p_{quarry}' = p_{quarry} + rand * precentage_i * (p_{quarry} - P_{falconer}) \quad (2)$$

$$P_{falconer}' = \begin{cases} rand * (2 * p_{quarry} - P_{falconer}) + p_{quarry} \\ \quad \text{if } (2 * p_{quarry} - P_{falconer}) < p_{quarry} \\ rand * p_{quarry} + (2 * p_{quarry} - P_{falconer}) \\ \quad \text{if } (2 * p_{quarry} - P_{falconer}) > p_{quarry} \end{cases} \quad (3)$$

$$P_{falconer}' = \begin{cases} rand * P_{falconer} + p_{quarry} & \text{if } P_{falconer} < p_{quarry} \\ rand * p_{quarry} + P_{falconer} & \text{if } P_{falconer} > p_{quarry} \end{cases} \quad (4)$$

The rand is used to generate a number between 0 and 1. The (3) is used to update the position of the falconer if it is in right or left section otherwise (4) is used for the same. The remaining females (which are moving towards territory) update their position according to (5).

$$P_{female\_lion}' = P_{female\_lion} + 2 * distance * \overrightarrow{R1} + U(-1,1) * \tan(\theta) * distance * \overrightarrow{R2} \quad (5)$$

Here, P_{female_lion}', P_{female_lion} is the updated and original position of the female lion respectively. The distance is the distance between the P_{female_lion}' and the point selected by tournament strategy. $\overrightarrow{R1}$ is the vector representing the start point, $\overrightarrow{R2}$ is selected such that $\overrightarrow{R1}.\overrightarrow{R2} = 0$ and $\|\overrightarrow{R2}\| = 1$. The θ varies uniformly between $-\pi/6$ to $\pi/6$ radian to widen the area of current solution. The $U(-1,1)$ is used for the generate -1 or 1 depending upon the direction of the resident male roaming in same or opposite direction of the pride territory respectively.

The tournament process enables the pride to change its size in each iteration by using (6):

$$Size_p^i = \max\left(2, ceil\left({}^{inol_i}/_{2}\right)\right) \quad i=1, 2, 3\dots\dots N \quad (6)$$

Where, $inol_i$ is the number of lion in pride i with improved fitness in previous iteration is given by equation (7).

$$inol_i = \sum_{j=1}^{n} sucess(j, iter, N) \quad i=1, 2,3\dots\dots N \quad (7)$$

Here the $sucess(j, iter, N)$ gives the success of lion j in group N at iteration iter denoted by (8)

$$sucess(j, iter, N) = \begin{cases} 1 & best_{j,N}^{iter} < best_{j,N}^{iter-1} \\ 0 & best_{j,N}^{iter} = best_{j,N}^{iter-1} \end{cases} \quad (8)$$

The nomad lion moves randomly in the search space (explorative search) used to avoid local optima. The movement of the ith nomad in the jth dimension is represented by (9)

$$nomad_{ij}^{lion'} = \begin{cases} nomad_{ij}^{lion} & \text{if } rand > pr_i \\ rand & \text{else} \end{cases} \quad (9)$$

Here, the pr_i is the probability generated for ith nomad given by (10)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$pr_i = 0.1 + \min\left(0.5, \frac{nomad_i - Best_{nomad}}{Best_{nomad}}\right)$$

i=1, 2, 3….number of nomads

(10)

where $nomad_i$, $Best_{nomad}$ is the cost of current position of ith nomad and best nomad respectively.

The lifestyle switching of the lion occurs due to different reasons. The strong male in the pride may beaten up other male of the same pride or any nomad may beat the male lion of the pride. The beaten up lion left the pride. The male as well as female lion get died with time. Some male lion switch their pride to takeover other pride. Some female also migrate from one pride to other. This may change the number of nomad and number of members in particular pride. The whole process is repeated to achieve the convergence until the stopping criteria reached [11].

4. PROPOSED WORK

This work designs a novel algorithm (CALOA- Classification assisted Lion Optimization Algorithm) for the big data classification based on LOA described in previous section. The input dataset having mXn dimension is classified by the proposed algorithm CALOA. Each instance of the dataset is considered a lion which is classified into pride (group of lion) by using the k-mean clustering with Schwarz criteria. The lion (instance) which doesn't belong to any pride (cluster) is marked as nomad. A pride is selected randomly and a few lion of the pride are marked as falconer. Then these falconers attack the quarry and update the position of the quarry and the falconer. The remaining members of the pride also update their position towards the convergence.

The phenomenon of the pride hunt is exhaustive search. The nomads update its position randomly which introduce the explorative search in the behaviour of algorithm. The lion switch their lifestyle i.e. nomad to pride or pride to nomad. Moreover, any lion can change their pride. This depends upon the ability of the lion (cost of fitness function). The process gets repeated until the convergence is achieved. The maximum number of iteration is chosen as the stopping criteria along with the convergence. The fitness function is opted as the classification accuracy explained in the next section.

The detailed process can be easily understood by the algorithm given next:

5. PROPOSED ALGORITHM

(CALOA- Classification assisted Lion Optimization Algorithm)

1. Input dataset having n attributes say (a1, a2, a3….an) and m instances.
2. Initiate no_of_pride=smallest value(default 2);
3. Apply K-means clustering to generate number of prides say $P_1, P_2, P_3, \ldots\ldots P_{no\_of\_prides}$.
4. For i=1:no_of_pride

   $SC = -2.\ln\hat{L} + k\ln(NOE(P_i))$

   Where NOE() is the function to calculate number of elements,

   k = the number of parameters to be estimated,

   $\hat{L}$ = The maximized value of the likelihood function given as $\hat{L} = p(x|\hat{\theta}, M)$

   where $\hat{\theta}$ are the parameter values that maximize the likelihood function.

   Apply K-mean on P_i Clusters for k=2 say generated prides are Pi1 and Pi2 .

   Calculate the SC for prides Pi1 and Pi2 by using

   $SC1 = -2.\ln\hat{L} + k\ln(NOE(P))$

   Here, the number of parameters gets doubled due to two prides.

   If SC>SC1 then

   no_of_pride=no_of_pride+1

   Pi=Pi1

   Pno_of_pride=Pi2

   i=i-1

   End if

   End for
5. Nomad=$\{\phi\}$
6. For i=1:m

   If $instance_i \notin P_j | \forall j \in \{1,2,3 \ldots\ldots no\_of\_pride\}$ then

   $Nomad = Nomad \cup instance_i$

   End if

   End for
7. falconer = random number between 1 and no_of_pride.
8. nv = random number between 1 to $NOE(P_{falconer})$.
9. Calculate P_{quarry}

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$P_{quarry} = \sum P_{falconer}(s_1, s_2, s_3, \dots \dots \dots s_{nv}) / no\_of\_falconer$$

10. Update position of quarry and falconer

$$p_{quarry}' = p_{quarry} + rand * precentage_i * (p_{quarry} - P_{falconer})$$

If falconer is in left or right section then use equation (3)

else

$$P_{falconer}' = \begin{cases} rand * P_{falconer} + p_{quarry} & \text{if } P_{falconer} < p_{quarry} \\ rand * p_{quarry} + P_{falconer} & \text{if } P_{falconer} > p_{quarry} \end{cases}$$

End if

11. Calculate female lion of the pride which are not falconer

$$P_{female\_lion} = P_{falconer} - P_{nv}$$

12. Update P_{female_lion}

$$P_{female\_lion}' = P_{female\_lion} + 2 * distance * \overrightarrow{R1} + U(-1,1) * \tan(\theta) * distance * \overrightarrow{R2}$$

The detail of the terms has been given in previous section.

13. Update the position of the nomads

$$nomad_i' = \begin{cases} nomad_i & \text{if } rand > pr_i \\ rand & \text{else} \end{cases}$$

14. Update pride size
15. if stopping criteria achieved then

Exit

else

Go to step 5

End if

The algorithm works for the following fitness function for the classification.

6. FITNESS FUNCTION

The cost of fitness function can be evaluated by using the (11):

$$Fitness_{cost} = c_1 * accuracy \quad (11)$$

The c1 is the constant taken as 0.8 on experimental basis. The accuracy can be easily evaluated by applying the j48 on each nomad to generate class for each instance.

The given optimization technique achieves the global optima better than the other existing metaheuristic techniques [11]. The better global optima assisted by classification technique results in better classification as compared to other existing techniques shown in next section.

7. RESULT AND DISCUSSION

The proposed calculation is dissected by adding the proposed calculation to the WEKA library on the Intel i5 @ 2.67 GHz utilizing the Eclipse IDE. The calculation is examined on the datasets having 400 occasions with 25 qualities and 32561 examples with 15 properties. The subtle element portrayal of these datasets is given in table 1.

Table1: Dataset Description

Information Type	Dataset 1 Medical	Dataset2 E-commerce
Number of attributes	25	15
Number of examples	400	32561
Number of classes	2	2
Trait Type	Numeric, Nominal	Numeric, Nominal
Reference	UCI repository[17]	UCI repository[17]

The proposed calculation of (CALOA) is executed on two datasets dataset1 - medical dataset and dataset2 - E-commerce dataset depicted in the table which are available on online UCI repository. Various parameters that are evaluated over dataset are accuracy, TP rate, FP rate, Precision and Recall. The portrayal of these parameters can be found in [15].

The table 2 demonstrates the execution correlation of proposed CALOA and RNFCA calculation on dataset1 and dataset2.

Table 2: Performance comparison of RNFCA and proposed CALOA *algorithm*

Datasets	Dataset1		Dataset2	
Parameters	RNFCA	CALOA	RNFCA	CALOA
Accuracy (%)	99.75	99.87	88.025	88.147
TP Rate	0.998	0.99	0.88	0.89
Fp Rate	0.002	0.001	0.228	0.213
Precision	0.998	0.999	0.88	0.89
Recall	0.998	0.999	0.873	0.885
F-Measure	0.998	0.999	0.871	0.892

The performance of the proposed CALOA, is also compared with our other existing CCSA algorithms [16], RNFCA [18] and the decision tree

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

i.e. J48. The figure 1 - 4 shows the comparison graph of the J48, CCSA, RNFCA and proposed CALOA over various parameters like accuracy, recall and F-measure on both datasets: dataset1 - medical dataset and dataset2 - E-commerce dataset.

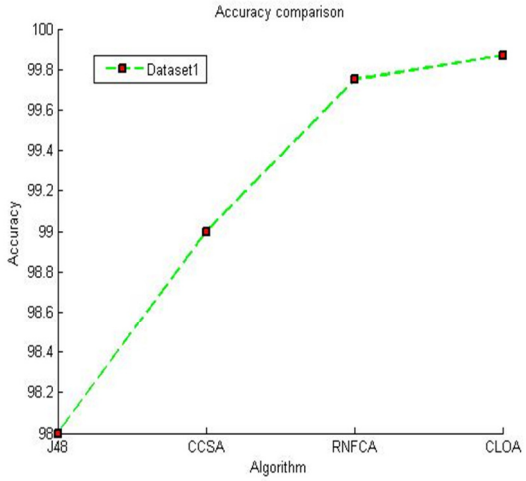

Figure 1: Comparison of Accuracy of J48, CCSA, RNFCA and Proposed CALOA *Algorithm on Dataset1*

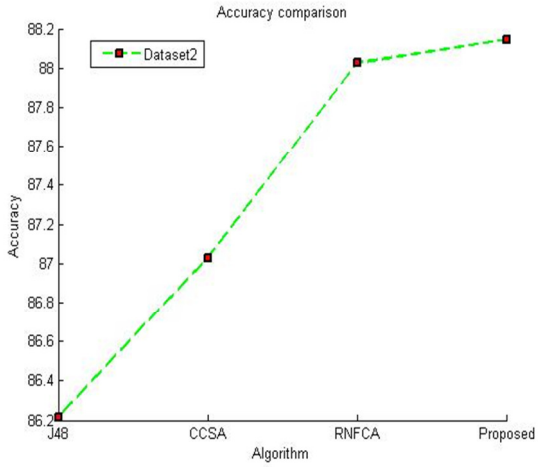

Figure 2. Comparison of Accuracy of J48, CCSA, RNFCA and Proposed Algorithm CALOA *on Dataset2*

The figure 1 and figure 2 demonstrates the correlation of the accuracy for the systems J48, CCSA, RNFCA and the proposed CALOA over dataset1- medical dataset and dataset2 – E-Commerce dataset determined in table 1. The analysis and examination determines that the proposed algorithm CALOA (Classification assisted Lion Optimization Algorithm) gives 0.12% improvement over RNFCA calculation which enhanced the exactness around 0.875% over the CCSA calculation which has 0.9% upgrade over the J48 calculation. The change in the accuracy results in upgraded mapping of components with right class. This shows significant change in accuracy using CALOA optimization technique over other existing algorithms.

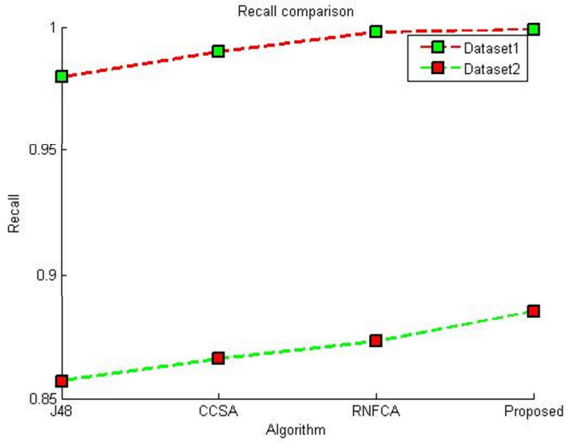

Figure 3. Recall comparison of the J48, CCSA, RNFCA and Proposed CALOA.

The parameter review indicates the importance of the yield. The upgrade of the review estimation of proposed CALOA calculation over the RNFCA, CCSA and J48 speaks to more pertinent information through proposed CALOA calculation.

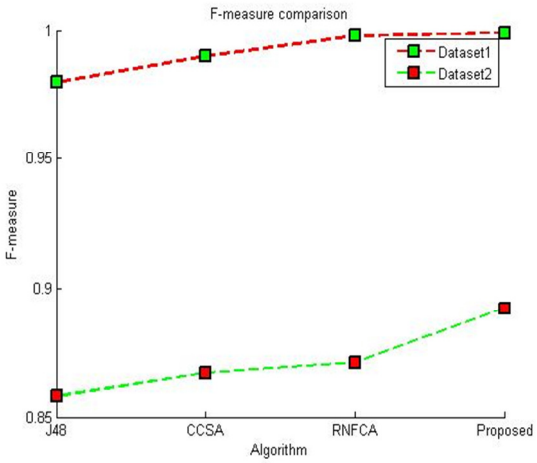

Figure 4. F-measure comparison graph of the J48, CCSA, RNFCA and Proposed CALOA algorithm

The F-measure is the consonant mean of the exactness and the review which measures the viability of the calculation. The proposed CALOA calculation displays better F-measure when

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

contrasted with RNFCA , CCSA and J48 calculation demonstrating centrality of the calculation.

The figure 5 compares the various algorithms of literature which incorporates the J48, ant colony optimization (ACO), Particle Swarm Optimization (PSO) and Differential Evolution (DE) algorithm with proposed algorithm CALOA on four datasets downloaded from the UCI repository [17]. These datasets have commonly being used to analyze the performance of the classification algorithms.

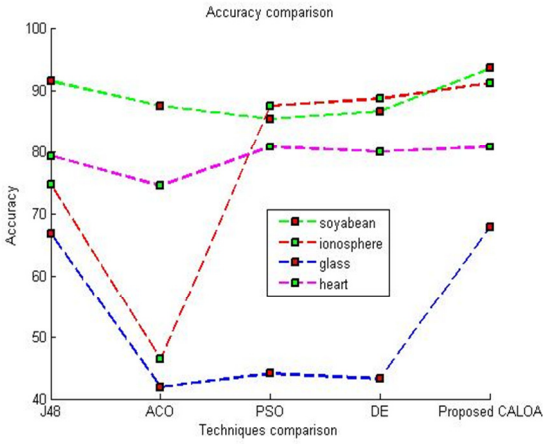

Figure 5: Comparison of Literature Techniques with proposed Algorithm CALOA

The figure 5 clearly shows the performance of the proposed algorithm CALOA is better than all the algorithms of literature i.e. J48, ACO, PSO and DE, while the performance of the proposed CALOA is minutely better than the RNFCA algorithm. It is due to the optimization nature of the proposed CALOA algorithm.

8. CONCLUSION

This paper has worked on big data classification by using lion optimization algorithm. The developed algorithm has used k-mean clustering with schwarz criteria for the initiation purpose. The prides formed by the algorithm is classified by using the J48 algorithm. The comparison of the algorithm with other cluster based classification algorithm by using the parameter like accuracy, recall etc. shows the significant improvement over other algorithms. The performance of the proposed algorithm CALOA is also compared and has been found better than other algorithms of literature i.e. J48, ACO, PSO and DE.

In future the algorithm can be extended for regression purpose. Moreover, the proposed CALOA algorithm can be analyzed for various applications domains to handle different types of big datasets.

REFERENCES:

[1] Jiban K Pal, "Usefulness and applications of datamining in extracting information from different perspectives", Annals of library and information studies vol. 58, March 2011, pp. 7-16

[2] Quinlan J. R., "Induction of decision tree", Machine Learning, vol. 1, pp. 81–106, Kluwer *Academic Publishers, Boston - Manufactured in the Netherland, 1986.*

[3] Quinlan J.R.,"Improved use of continuous attributes in c4.5", *Journal of Artificial Intelligence Research*, Vol 4, pp 77-90, 1996.

[4] Kohavi R., Quinlan J. R., "C5.1.3 Decision-tree discovery", 1999.

[5] Breiman L., Friedman J. H.,"Classification and Regression Trees"', *Wadsworth, Belmont, CA Chapman & Hall, New York*, 1984.

[6] Holland JH., "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence", *U Michigan Press*; 1975.

[7] Qinghai Bai, "Analysis of particle swarm optimization algorithm", *Computer and information Science,* Vol 3 No. 1Feb, 2010.

[8] Rafael S. Parpinelli, Heitor S. Lopes, and Alex A. Freitas,"Data Mining with an Ant Colony Optimization Algorithm", *IEEE Transactions on Evolutionary Computation* 6(4), pp 321 – 332.

[9] Esmat Rashedi, Hossein Nezamabadi-pour, Saeid Saryazdi ,"A Gravitational Search Algorithm ", *Elsvier Information Sciences* 179, 2009, 2232–2248.

[10] Yang X-S. ,"A new metaheuristic bat-inspired algorithm", *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010).* Springer; 65–74.

[11] Yazdani, M. and Jolai, F., "Lion Optimization Algorithm (LOA): A nature-inspired metaheuristic algorithm", Journal *of Computational Design and Engineering,* 3(1), pp.24-36, 2016..

[12] Yu-JunZheng, "Water wave optimization: Anewnature-inspired metaheuristic", *Computers OperationsResearch*55, 2015, 1–11.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[13] Xiaofen Lu and Ke Tang, Xin Yao,Classification-Assisted Differential Evolution for Computationally Expensive Problems, *IEEE Congress on Evolutionary Computation (CEC),* 2011

[14] H.R., Tizhoosh, "Opposition-based learning: a new scheme for machine intelligence", *Proceedings of the CIMCA/IAWTIC*, 2005.

[15] Navneet and Nasib Singh Gill, "Algorithm for Producing Compact Decision Trees for Enhancing Classification Accuracy in Fertilizer Recommendation of Soil", *International Journal of Computer Applications,* 98(2), 2014, pp 8-14.

[16] Navneet and Nasib Singh Gill, "Classification using the compact rule generation", *Oriental Journal of Computer Sc. and Technology*,8(1) ,2015,pp 49-58.

[17] Lichman, M.,"UCI Machine Learning Repository ", [http://archive.ics.uci.edu/ml], *Irvine, CA: University of California, School of Information and Computer Science*, 2013.

[18] Navneet and Nasib Singh Gill, "An Optimized Algorithm for Big Data Classification using Neuro Fuzzy Approach", *Indian journal of science & technology,* Vol 9(28), 2016.