International Journal of Engineering & Technology, 7 (3.29) (2018) 554-557

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Binary gravitational search algorithm (BGSA) for solving feature selection problem

NageswaraRao Banoth [1] *, Suresh Dara y [1, 2], M. Jagadeeshwara Reddy [2], R. P. Singh [3]

[1] Department of Computer Science and Engineering, JBIET, Hyderabad, Telangana-500075, India
[2] Department of Computer Science and Engineering, B.V.Raju Institute of Technology, Narsapur, Telangana-502313 India
[3] Sri Satya Sai University of Technology & Medical Sciences , Sehore , M.P -466001, India
**Corresponding author E-mail: nageswararao.bano@gmail.com*

Abstract

In previous years, different Lateral thinking optimization techniques have been developed based on evolutionary computation. Many of these methods are inspired by spill out behaviors in nature. In this Paper, a new optimization algorithm based on the law of gravity and mass interactions named as Gravitational Search Algorithm (GSA) is discussed for solving feature selection. In GSA, the searcher agents are a collection of masses which will interact with each other based on the law of motion and Newtonian gravity which gives the binary evolutionary optimized high performance. The detailed feature selection has been discussed in this paper and The GSA method has been compared with some wellknown optimized search methods such as GA (Genetic Algorithm), PSO (Particle Swarm Optimization).

Keywords: *Gravitational Search Algorithm; Evolutionary Algorithms; Law of Gravity; Optimization.*

1. Introduction

Feature Selection (FS) is the combination of the algorithms for taking the set of features which are subsets Suppose initialize the set with N Number of features in set F:

$$F = x_1, x_2, x_3, x_4, \ldots, x_n \tag{1}$$

Then the subsets of the features are:

$$F^1 \subset F = x_1^1, x_2^2, x_3^3, \ldots, x_m^1 \tag{2}$$

The number of possible subset for N Number of sets
is 2^N.
FS has turned into an indispensable piece of information characterization issues, inferable from the substantial number of properties that the true datasets contain. It has been investigated utilizing customary techniques for different grouping problems. Some of these analysts have played out their investigations on therapeutic datasets also. They have utilized conventional channel and wrapper strategies to look at restorative datasets.
Feature Selection is a preprocessing method for compelling information examination. The motivation behind component choice is the determination of ideal subsets, which are important and adequate for tackling the issue. Feature determination enhances the prescient exactness of calculations by lessening the dimensionality, evacuating insignificant Features and diminishing the measure of information required for the learning procedure. This should be possible in light of the fact that not all accessible Features are significant for the order procedure. As of late, include choice has been effectively utilized to ade-quately take care of order issue in different territo-ries, for example, design acknowledgment, information mining , and sight and sound data recovery and different regions where Feature determination can be connected to Which will not possible for n number of sets so we use algorithms for them.
Feature selection is active area in different fields such as data processing, data mining, machine learn-ing, classification problems. Until now, large number of methods for fs have been introduced and reported in literatures. Based on the algorithm selection and the model building the methods are categorized into four groups:

1) Filter method
2) Wrapper method
3) Embedded method or hybrid method
4) Ensemble method [1]

Very recently many evolutionary computation al-gorithms have been proposed for feature selection. Genetic Algorithm based feature selection [2], using Particle Swarm Optimization [3], Firefly algorithm based FS [4] etc.

2. Preliminaries

In this section, some of required fundamentals have discussed like GSA and FS with GSA to understand our proposed work.

2.1. GSA

The Gravitational search algorithm [5] is used for solving the feature selection problem which uses the Newtonian laws and agents. In GSA the items will be specialists and the execution will be estimated by their mass. Every one of the items pulls in each other protest by GF, and Gravitational power which ac-tivity the development of the articles all around to-wards different items with heaviest mass. The heavier masses which gives best aftereffect of the issue.

In the Gravitational search calculation, each spe-cialist has four Attributes: position of the opera-tor, its inertial mass, dynamic gravitational mass and aloof gravitational mass. Position of the mass com-pares to the issue arrangement, its inertial and gravitational masses are dictated by utilizing the fitness work.

Here each mass can take a gander at the proficiency of different mass, with the end goal that the GF is an Information exchanging instrument.

The drive from the closest specialists follows up on the mass, the operator can watch the region around it.

The heavier mass has more fascination width, so the immense viability of fascination. Thus, the mass with more viable execution will have the most gravi-tational mass. As the above proclamation says so the specialist will go to the better operator.

Gravitational pursuit calculation is memory less however it works like the calculation with memory. We used binary GSA algorithm in our work [6].

2.2. FS with GSA

The element choice strategy in view of GSA and OPF [7]. Here, principle topic is to utilize the OPF exact-ness as the fitness esteem over the arrangement of assessment to amplified bu GSA. So, eve-ry last op-erator is the likely arrangement in the dimensional-ity space, where the 1 esteem demonstrates that element is chosen do the new informational index and 0 esteem demonstrates that com-ponent isn't selected. The FS calculation is joins improvement of the GSA [5] with the OPF speed [5] classifier. Here [8] uti-lized the accuracy of the OPF as capacity to course GSA into looking through the better arrangements. The arrangement of the vector came about by GSA is more trustable as the majority increments. Thus, there is require in classifier speedier for preparing all conceivable subset of Feature assigned by every molecule position. The calculation taken donot stop the OPF as classifier, utilized for efficacy of train-ing.It doesnot enhancement issue in parameter, for example, neural networks.Algorithm [8] is utilized to include choice. The power evaluation is finished.

For the Evaluation of the information of therapeu-tic for expectation of sickness requires the procedures of Feature Selection which are effective, the informa-tion have the tremendous number of the Features. Research have done utilizing EC (Evolutionary cal-culation, for example, GA [9], PSO [10] for include determination and discovered them speedier than the ordinary procedures. So [11] utilized the very nearly another system in the field of restorative called GSA for the choice of Features in the datasets. The tech-nique for wrapper based is utilized, brushing the GSA and K-closest expels the undesirable information by normal of 66% enhancing the forecast exactness.

The fitness function is used for better position is:

$$Fitness = \alpha \times \gamma + \beta \times (T - S)/S \tag{3}$$

This experiments are done number of times to deduct the arbitrary factors.

The KNN is utilized for the reason for order. The grouping effec-tiveness is estimated on the

$$accuracy = \frac{TPO + TNE}{TPO + TNE + FPO + FNE} \tag{4}$$

keeping the overfitting the approval is finished. The lessening is better i.e.., 64.61% and the ac-curacy has discreetly expanded from the scope of 64.81% to 82.96% in the dataset of heart

The Dermatology set diminished to 57.64% and ef-cieny of classifier is 87.14% to 96.71%. In the bo-som growth the higher esteem is acquired 77.77% and slight decrease of the precision 97.14% to 95.7%.

The component determination is utilized with the GSA which is modified [12].The altered GSA direct picewise guide of turbulent to build the species decent variety and the quadratic consecutive quickening of programming of nearby investigation. The worked is to ad lib and enhance the choice by the Modified GSA. Distinc-tive examinations and comparisions [11] are done. The framework has better execution and precision has higher which is accom-plished contrasted with accessible dataset and different frameworks. The Behjat A.R led the segment of Feature in the Security frame-work as interruption framework to con-trol PCs which are joins frameworks this framework.

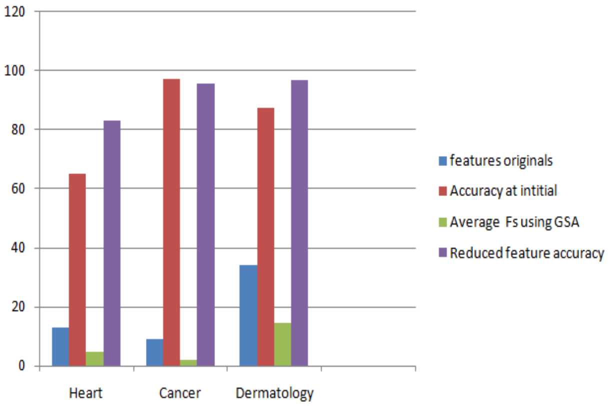


Fig. 1: Reduction Rate.

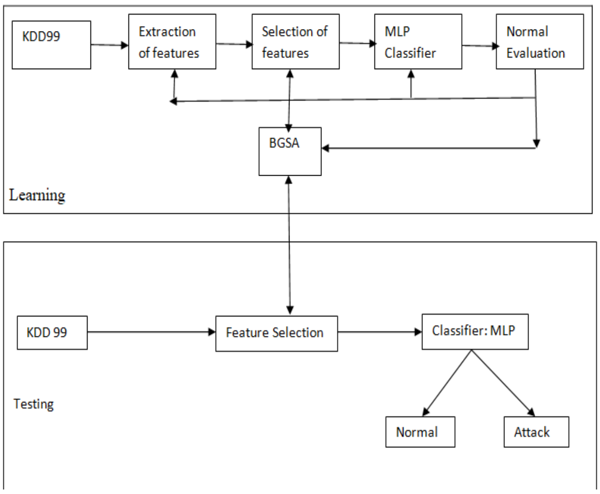


Fig. 2: Feature Subset Selection Using Binary Gravitational Search Algo-rithm.

has assumed an unmistakable part to improvise the ase rate at its least the diverse strategies are uti-lized The BGSA [6] as the determination for Features lessens the non-needed Features in KDD 99 recogni-tion of interruption framework and extemporizes the multilayer execution with most minimal calculation cost the expansion has gone to about 100%

The KDD framework is utilized with the end goal that the intrusion and class 2 is utilized for better outcomes

We Present the Binary Gravitational search algo-rithm for solving the feature selection problem where the number of redundant and irrelevant features are more in huge datasets such that to reduce the redun-dant features and to obtain the binary vector values for computational purpose for the fitness function we take and gives the best result There are numer-ous Evolutionary enhancement issues, for example, include determination and dimensionality diminish-ment [13] in which is to do the arrangements as double vectors. Also, issues in the dimensional space are taken in the double space, as well. The best arrange-ment is to demonstrate the digits entire numbers as opposed to in double digits.

The essential ideas of GSA are unquestionably al-tered in the Binary GSA. Here in the parallel condi-tion, each measurement can take just 0 or 1. Trav-eling through the measurement implies experiencing the 0 or 1.

Principally after the main emphasis of the calcu-lation in the updation of the speed here we utilize the

$$f(x) = \begin{cases} p_e^f(t+1) = \ (p_e^f(t)), & r_j < S(v_e^f(t+1)) \\ p_e^f(t+|1) = p_e^f(t), & else \end{cases}$$

To accomplish the twofold form for the following cycle. The principle perception between the GSA and BGSA is that the position refreshing is finished utiliz-ing the twofold form 0 or 1 esteems. It is finished by the speed of mass and considering the esteem if the position which is not as much as the arbitrary esteem ought to be taken as 0 or in the event that it is more than the irregular esteem it ought to be considered as

1) A little estimation of the speed and position must give the little likelihood of position evolving. An ex-tensive estimation of speed should give vast likelihood of the adjustment in the mass position from its past position. At the end of the day, minimal estimation of the speed will give the great mass position which ought not be changed (Considering the minimum es-teem is 0) it is noticed the esteem is ascertained ham-ming distance [14].

In a same way of BPSO, the speed is considered in BGSA as a likelihood. Be that as it may, in GSA, a position refreshing means an exchanging between the two conceivable 452 M. Sarhani et al. values. As it were, it demonstrates the likelihood of changing the estimation of $x_i^j\ (t)$ from "0" to "1" and the other way around. Likewise, the change is finished utilizing the tanh work rather than sigmoid capacity as characterized in Eq.:

$$f(x) = \begin{cases} p_e^f(t+1) = \ 1, & r_j < S(v_e^f(t+1)) \\ p_e^f(t+1) = 0, & else \end{cases}$$

3. Proposed work

3.1. Fitness function

The entire preprocessing approach and the fitness function is consider as it is available in literature [14], [15] and applied in our current work. At the point when the preprocessing was finished, we got the de-creased yet high dimensional qualification table as a yield like the one as appeared in Table I. The com-ponent choice should be possible by BGSA utilizing the accompanying target work. We proposed a fit-ness function, which incorporates two sub capacities (F1andF2). Where F1 discovers number of Features (i.e. number of 1's), F2 chooses the degree to which the element can perceive among the protest's sets. The used fitness function is:

$$Fit = \alpha_1 F_1(\vec{v}) + \alpha_2 F_2(\vec{v}) \tag{5}$$

The proposed feature selection algorithm using BGSA shown in Algorithm (1).

3.2. Datasets

We have implemented the BGSA Algorithm to find minimal feature subsets on Cancer datasets. The de-tails of taken publicly available datasets are as shown in Table 1.

Table 1: Details of the Cancer Datasets

Datasets	Total	Reduced	Classes	Samples		
	Features			Total	Train	Test
Colon	2000	1102	Colon cancer	40	20	20
			Normal	22	11	11
Lymphoma	4026	1867	Other type	54	27	27
			B-cell	42	21	21
Leukemia	7129	3783	ALL	47	27	20
			AML	25	11	14

4. Results & comparisons

Table 2 is showing undertook k-nearest neighbors (kNN) classier, for different k-values (1, 3, 5 and 7, an odd number to eredicate ties) and the respective recognition values on test set. The correct classification is 80:65%; 83:92% and 81:48% for these three datasets. Note that when k = 1, all datasets give 100% correct classification. After Preprocessing

Table 2: Performance on Three Data Sets Using K-NN Classifer

Dataset	Popu-lation size	Selected subset features	k-nearest neighbors classification (%) on test set			
			$k=1$	$k=3$	$k=5$	$k=7$
Colon:	10	3	100	80.65	79.65	76.20
# Genes 2000	20	8	100	75.42	75.20	65.75
Reduce to 1102	30	11	100	75.42	63.52	63.52
	50	12	100	80.65	74.20	74.20
Lymphoma:	10	14	100	83.92	83.75	80.58
# Genes 4026	20	12	100	79.59	82.25	76.09
Reduce to 1867	30	19	100	79.59	83.42	82.25
	50	15	100	74.17	80.25	75.84
Leukemia:	10	13	100	81.22	80.85	75.32
# Genes 7129	20	11	100	76.85	79.58	77.32
Reduce to 3783	30	18	100	72.11	75.22	81.85
	50	17	100	81.48	78.22	80.48

Table 3: Comparative Performance with GA-FS Algorithm

Data	Method	Classifier Method		
		DS	LibSVM	Beyesian
Colon	GA-FS	74.24	78.9	70.10
	BGSA	74.12	**80.88**	79.44
Lymphoma	GA-FS[16]	75.9	77.4	70.0
	BGSA	**73.88**	**75.10**	**71.98**
Leukemia	GA-FS	75.8	79.4	77.2
	BGSA	70.11	**80.50**	**72.52**

Classifier Method

Asha Gowda Karegowda [17] proposed channel, GA with FS as subset-assessing component has been explored different avenues regarding therapeu-tic datasets. While GA guarantees worldwide in-quiry, CFS brings about lessened component subset. What's more CFS is very connected with the class have low inter-correlation. The test comes about un-mistakably show that the channel GA FS enhances order precision of SVM and classifiers for all the ther-apeutic dataset. The Bayesian classifier execution

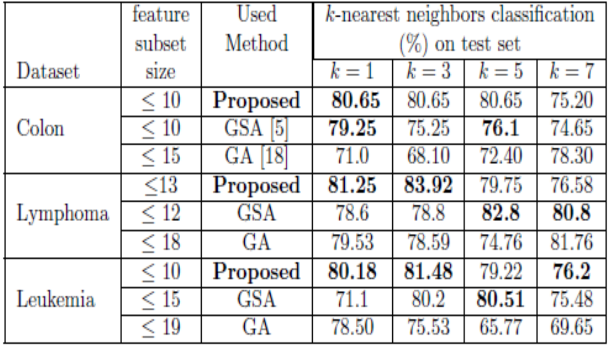
```
Algorithm 1 The BGSA Algorithm for Feature Selection
Step: 1 Initialize population randomly
Step: 2 Evaluate the fitness for each agent by using
        Fitness function 5
Step: 3 Update the Gravity, best and worst of the
        population.
Step: 4 Calculate Mass and acceleration for each
        agent
Step: 5 Update velocity and position
Step: 6 Return to Step-2 if not meet exit condition,
        otherwise Step-7
Step: 7 Return best Subset
```

donot enhance to much obviously, neither did not de-crease with less number of applicable sources of info gave by GA CFS .

But the proposed BGSA has demonstrated ensuing increment 3 in the classifier technique for SVM to 80.88% and separately with the other two datasets to 75.10% and 80.50%. At the point when con-trasted with the taken BGSA proposed calculation.

Table 4 is showing that proposed one has the best classification factor with less than ten features sub-set size when the value of k = 1 compared to GSA 75.25% and 78.30% for GA which is 80.65% for the Colon Dataset. For, the Lymphoma Dataset the GSA has the classification accuracy of 82.8% and GA has the 81.76% but BGSA has proven that it is best which is 81.25% Another Dataset, Leukemia Also shown the best classification rate in the BGSA 81.48% while when compared to the GSA and GA classification rate 80.51% and 78.50%. From the above experimen-tal results, it is clear that our proposed BGSA al-gorithm shows better and comparative performance with the existing ones on bench mark high dimen-sional datasets.

Table 4: Comparative Performance between BGSA, GSA and GA on Three Datasets Using K-NN Classifier



Dataset	feature subset size	Used Method	k-nearest neighbors classification (%) on test set			
			$k = 1$	$k = 3$	$k = 5$	$k = 7$
Colon	≤ 10	**Proposed**	**80.65**	80.65	80.65	75.20
	≤ 10	GSA [5]	**79.25**	75.25	**76.1**	74.65
	≤ 15	GA [18]	71.0	68.10	72.40	78.30
Lymphoma	≤ 13	**Proposed**	**81.25**	**83.92**	79.75	76.58
	≤ 12	GSA	78.6	78.8	**82.8**	**80.8**
	≤ 18	GA	79.53	78.59	74.76	81.76
Leukemia	≤ 10	**Proposed**	**80.18**	**81.48**	79.22	**76.2**
	≤ 15	GSA	71.1	80.2	**80.51**	75.48
	≤ 19	GA	78.50	75.53	65.77	69.65

5. Conclusion

We proposed a BGSA Algorithm for discovering Fea-ture subsets from high dimensional quality Bioinfor-matics information. At first, the information has been preprocessed and discretized utilizing a quick heuristic strategy and after that the parallel refine-ment table is created. The proposed BGSA is then connected to recognize discriminative and critical qualities from high dimensional quality Bioinformat-ics datasets. The parameters of BGSA with various populace measure are additionally explored for pro-mote change of the outcomes. The clashing necessity of the component choice is, to choose negligible ele-ment subsets with same or higher grouping exactness as the entire capabilities. Here, these objec-tives were accomplished through appropriate joining of two fit-ness capacities. The execution of the proposed technique and the cur-rent strategies were looked at by utilizing the pre-scient precision of standard classifiers. An essen-tial finding is that the proposed include determina-tion calculation is appeared to be more power-ful in choosing vital Features from high dimensional qual-ity Bio-informatics datasets. The outcomes have been tried on three benchmark Cancer datasets and pre-sented results and compari-sons to prove out perfor-mance of proposed work.

References

[1] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinfor-matics (TCBB), 9(4):1106{1119, 2012.

[2] Emina Alickovi and Abdulhamit Subasi. Breast cancer diagnosis using ga feature selection and rotation forest. Neural Computing and Applica-tions, 28(4):753{763, 2017.

[3] Yong Zhang, Dun-wei Gong, and Jian Cheng. Multi-objective par-ticle swarm optimization ap-proach for cost-based feature selection in clas-sification. IEEE/ACM Transactions on Com-putational Biol-ogy and Bioinformatics (TCBB), 14(1):64{75, 2017.

[4] Yong Zhang, Xian-fang Song, and Dun-wei Gong. A return-cost-based binary firefly algo-rithm for feature selection. Information Sci-ences, 418:561{574, 2017.

[5] Esmat Rashedi, Hossein Nezamabadi-Pour, and Saeid Saryazdi. Gsa: a gravitational search algo-rithm. Information sciences, 179(13):2232{2248, 2009.

[6] Esmat Rashedi, Hossein Nezamabadi-Pour, and Saeid Saryazdi. Bgsa: binary gravitational search algorithm. Natural Computing, 9(3):727{745, 2010.

[7] Joao P Papa, Alexandre X Falcao, and Celso TN Suzuki. Super-vised pattern classification based on optimum-path forest. Interna-tional Journal of Imaging Systems and Technology, 19(2):120{131, 2009.

[8] Suresh Dara and Haider Banka. A binary pso feature selection algo-rithm for gene expression data. In Advances in Communication and Com-puting Technologies (ICACACT), 2014 Interna-tional Con-ference on, pages 1{6. IEEE, 2014.

[9] Kalyanmoy Deb. An introduction to genetic al-gorithms. Sadhana, 24(4-5):293{315, 1999.

[10] James Kennedy. Particle swarm optimization. In Encyclopedia of machine learning, pages 760{766. Springer, 2011.

[11] Sushama Nagpal, Sanchit Arora, Sangeeta Dey, et al. Feature selec-tion using gravitational search algorithm for biomedical data. Pro-cedia Com-puter Science, 115:258{265, 2017.

[12] Xiaohong Han, Xiaoming Chang, Long Quan, Xiaoyan Xiong, Jingxia Li, Zhaoxia Zhang, and Yi Liu. Feature subset selection by gravitational search algorithm optimization. Information Sci-ences, 281:128{146, 2014.

[13] Dingcheng Feng, Feng Chen, and Wenli Xu. Su-pervised feature subset selection with ordinal op-timization. Knowledge-Based Sys-tems, 56:123{140, 2014.

[14] Haider Banka and Suresh Dara. A hamming dis-tance based binary particle swarm optimization (hdbpso) algorithm for high dimen-sional feature selection, classification and validation. Pattern Recog-nition Letters, 52:94{100, 2015.

[15] Suresh Dara Chandra Sekhara Rao Annavarapu and Haider Banka. Cancer microarray data fea-ture selection using multiobjective bi-nary particle swarm optimization algorithm. EXCLI journal, 15:460, 2016.

[16] Mehmet Fatih Akay. Support vector machines combined with fea-ture selection for breast can-cer diagnosis. Expert systems with ap-plications, 36(2):3240{3247, 2009.

[17] Asha Gowda Karegowda, MA Jayaram, and AS Manjunath. Fea-ture subset selection using cascaded ga & cfs: An filter approach in super-vised learning. International Journal of Com-puter Applica-tions, 23(2) one {10, 2011.

[18] Derrick Joel Zwickl. Genetic algorithm ap-proaches for the phylo-genetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, 2006.