ISSN 2007-9737

Classification of Complex Diseases using an Improved Binary Cuckoo Search and Conditional Mutual Information Maximization

Manu Phogat, Dharmender Kumar

Guru Jambheshwar University of Science and Technology, India

{kunjean4181, dharmindia24}@gmail.com

Abstract. With the advancement of various computational techniques, there is an exponential growth in genomic data. To analyze such huge amount of data, there is necessity of efficient machine learning techniques. The genomic data usually suffers from "curse of dimensionality" problem, having large number of n (features) and small number of p (samples), which makes classification task very complex. In the present study, a new intelligent hybrid method based on CMIM (conditional mutual information maximization) and novel IBCS (Improved binary cuckoo search) is used for classifying various complex diseases. The CMIM is used to deal with dimensionality problem and IBCS is to select most informative features. Generally, the standard BCS (binary cuckoo search algorithm) is used for feature selection but it has problems like low optimization accuracy and low localized searching. The IBCS overcome the shortcomings of BCS, and improved the classification accuracy by choosing best informative feature subset. The proposed technique applied on five different SNPs dataset which are publically available on NCBI GEO. The proposed model attains high classification accuracy and outperformed other feature selection techniques. The IBCS was also compared with other metaheuristics algorithms such as Binary GA, Binary PSO and Binary ACO, and the result shows that it has better classification accuracy.

Keywords. Metaheuristic, CMIM, IBCS, feature selection, classification, SNP

1 Introduction

The Metaheuristics algorithms, which are usually nature inspired algorithms, are quite popular in recent times because of their high efficiency in solving various challenging optimization problems. The nature-inspired algorithms are often developed by mimicking the behavior of some species in nature (e.g. humans, insects, birds).

Recently the metaheuristics algorithms are widely applied in the field of genomic data for selecting the suitable genetic marker (genes) to classify various complex diseases [5].

The Genome Wide Association Studies (GWAS) shows that a variety of diseases are characterized by SNP (single nucleotide polymorphism) [18]. The SNPs are type of genetic variation with a substitution of a single nucleotide in the DNA sequence of human genome [7]. The SNPs datasets are recently used for the categorization of various complex diseases. The SNP datasets are known for their high dimensionality and also contains huge level of noisy data. The dataset are basically consist of large number of features(SNPs) and a very small sample size, in machine learning it is considered to be "curse of dimensionality" problem and for this kind of dataset it is not easy to establish an efficient classifier [12].

To classify the complex diseases with SNPs datasets, one has to select the highly discriminative features (SNPs) and for that task a robust feature selection technique is used in computational intelligence [6]. When the high dimensional genomic data is characterized with traditional classifier for the diagnoses of a disease, a very low accuracy is obtained.

The main objective of feature selection in SNPs datasets is to minimize the features and maximize the classification performance. The metaheuristics algorithms have capability to efficiently search the entire set of features to find out the most informative features. In optimization

techniques the process of feature selection comes to end when the objective function reaches near to optimum solution. Many metaheuristics methods are successfully applied to genomic datasets for feature selection [8].

The SNPs are excellent genetic markers for many complex diseases, so our aim is to find out the interactions and relationships between SNPs to enhance the classification performance. Due to the high dimensionality and large feature space of SNPs dataset, the combination of intelligent feature selection techniques are used for the prognosis of disease and higher classification accuracy. Anekboon *et al.* [2] used a hybrid FS method by using three techniques, CBFS used as a filter and K-NN and ANN are used in the wrapper phase.

In this paper, we used the CMIM algorithm as a filter technique to reduce the size of large feature space by selecting a small feature subset on the basis of relevancy and redundancy. The chosen subset is then provided as input to proposed IBCS (Improved binary cuckoo search algorithm) to find out the most informative feature which are deeply related to disease.

The IBSC algorithm is an improved version of binary cuckoo search algorithm with having two main objectives.1) to retain only the useful features from the feature subset. 2) Maximize the predictive accuracy. The proposed technique is compared with many FS techniques, which are used on SNPs datasets, and also some metaheuristics algorithms, which are newly applied on the selected SNPs datasets.

The workflow of the paper is designed as follows: Section 2 describes the proposed technique, Improved Binary cuckoo search and CMIM algorithm. Section 3 discusses the methodology and result and finally section 4 describes the conclusion.

CMIM Algorithm

The CMIM is a filter FS technique proposed by Fleuret [9] in 2003-2004. The CMIM algorithm selects features on the basis of conditional mutual information. It deals with independence and individual strength of a new feature by measuring it with the feature that is already taken.

Let's say a feature M' is only good if $I(y,\frac{M'}{M})$ for each M already picked. $y \rightarrow target \quad I \rightarrow Information, \quad y \rightarrow$ Boolean random variable for class. It stands that M' is only good if it possesses the information about y, and the information has not been shown by already picked up feature M. The CMIM criterion is expressed in equation 1:

$$CMIM(M_n) = min_{J \in S} I\left(M_{i,} \frac{Y}{M_j}\right), \qquad (1)$$

M_n= feature relevant to target y,

$S =$ Already selected feature.

The above equation explains that a higher value of feature(M_n), means it is relevant to target y and strongly complementary to another marked feature M_j where$J \in S$.

CSA (Cuckoo Search Algorithm)

The CS algorithm lies in the category of swarm intelligence, it is an optimization algorithm motivated by obligate brooding behavior of cuckoo bird in laying their eggs in other birds nest [20]. The cuckoo bird has a very exotic deception strategy that they replace one egg of the host bird with their own egg. The color and pattern of cuckoo egg resembles to the host eggs. Some of specific species of cuckoo bird have evolved in way that they have become specialized in mimicking the color and pattern of the eggs of few specific host eggs [11].

This strategy helps the cuckoo egg hatch slightly ahead before the host egg and because of their excellent mimicking pattern; the host throw out its own egg from the nest blindly, this process will increase the survival of cuckoo chicks. The idea of using cuckoo bird strategy in optimization was proposed by Yang and Deb [23] in 2009-2010 and can be used for numerous optimization problems.

The basic cuckoo search algorithm follows three simple rules:

R1: The cuckoo bird lays their egg in randomly chosen nests.

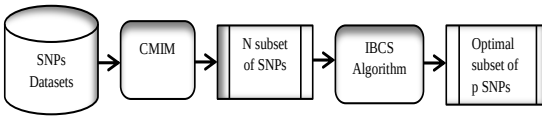

Fig. 1. Proposed Methodology

R2: The numbers of available nests are fixed, and the nest with the highest quality of eggs forms the next generation.

R3: When the host bird identifies the cuckoo's egg it has two choices either discard the alien egg or build a new nest for itself.

In the optimization scenario, the host nest is a possible solution for the problem. The first step of the algorithm is to randomly initialize the nest for each iteration.

The second step is to use Levy flight technique shown in Equation 1 and 2 for global random walk and update the position of the nests [22, 21]:

$$X_i^j(t) = X_i^j(t-1) + \alpha \oplus L(s,\lambda)\,, \tag{2}$$

$$L(s,\lambda) = \frac{\lambda.\Gamma(\lambda).sin(\lambda)}{\pi}.\frac{1}{s^{1+\lambda}}, \quad s \gg s_0 > 0\,. \tag{3}$$

X_i^j Represents for the j_{th} egg at nest$i$, $i = 1, 2, 3, \ldots, m$ and $j = 1, 2, \ldots . d$, s = step size α >o is a step size scaling factor.

BCSA (Binary Cuckoo Search Algorithm)

The binary version of cuckoo search is known as Binary Cuckoo Search algorithm (BCS), which is used for various feature selection techniques [19]. The main purpose is to collaborate a set of binary values for each nest that represents, whether a feature would reside to the conclusive set of features or not, and the function that has to be maximized is being provided by an organized classifier's accuracy.

To have solution in discrete form, a limit is applied for valuing the dimensions (eggs) by setting up the binary vectors UB=1(upper bound) and LB=0 (lower bound) in BCS. The solution generated and updated are having values between LB and UB. To choose a particular feature a binary vector is applied, where '1' stands for selected feature and '0; for unselected [16].

The binary values with in the Boolean lattice provide by equation 3 and 4:

$$S\left(x_i^j(t)\right) = \frac{1}{1+e^{-x_i^j(t)}}, \tag{4}$$

$$x_i^j(t+1) = \begin{cases} 1 & if\ s(x_i^j(t) > \sigma \\ 0 & otherwise \end{cases}, \tag{5}$$

where $\in (0,1)$, σ is a Boolean boundary and $s(x_i^j(t)$ is a sigmoid function.

The cuckoo search use the Levy flights search strategy to explore the search space in a straight path with randomly 90 degree turns [20], also CS mostly depend on random walk search to easily jump from one region to another without thoroughly exploring each cuckoo's nest.

Therefore CS has disadvantages like; low optimization performance, slow convergence rate and weak local searching. To overcome these shortcomings, we propose an IBCS (Improved Binary Cuckoo Search) algorithm.

2 Proposed Methodology

The proposed methodology is used to classify the SNPs datasets for various complex diseases as shown in Fig 1.

Firstly the filter method CMIM is used to reduce the large feature space into a small feature subset. CMIM technique discards the redundant and uninformative features. A fast implementation of CMIM algorithm is applied instead of standard algorithm.

It takes the feature score at the time of selection process and define CMI (conditional mutual information) for the features that are rich in info and less in redundancy. The top N features which maximize the mutual information among them and also with the target class, will be picked iteratively.

After the selection of N features by CMIM, the selected features are provided as an input dataset to IBCS algorithm to choose most relevant subset of features that can increase the classification accuracy to predict the complex diseases.

Algorithm 1. Pseudocode of proposed technique

Input: Total samples (X), Class labels (Y) Max_features, Final_features, initial_feature set (S)
Output: final selected feature set (N)

1 Set *Selected features* = null
2 for each features s_i in *S* do
3 Measure *MIi,* mutual information
4 Set *Pi= MIi*
5 Set *PUi* = 0 previously used
6 end
7 for *k* = 1 to *Max_Features*
8 do
9 Set $score_k$ = 0
10 for every features *si* in S
11 do while *Pi* > $score_k$ AND *PUi* < *k* - 1 **do**
12 Set *PUi =PUi* + 1
13 Compute CMI_{ik} between s_k and s_i
14 Set *Pi* = *min*(*Pi*; CMI_{ik})
15 end while
16 if *Pi* > $score_k$ **then**
17 Set $score_k$ = *Pi*
18 *F_SelectedFeatures=F_Selected Features* ∪ {s_i}
19 end
20 end
21 end
22 Set *N* = *finalSelectedFeatures*

IBCS Algorithm

*Dataset N, *Maxeval* (Stop criterion), No of feature dimensions *d,* No of nests *n,* U_b=upper bound, L_b= lower bound, A →Max levy step size
*Output parameters *B_fitness, B_Nest*

23 for *all nest* n_i (*i=1,. . . ,m*) do
24 for *all dimension j* (*i=1,. . . ,m*) do
25 $x_i^j(0)$ Random *[*L_b, U_b*]*
26 end
27 Convert n_i to binary vector using Eq. 4 & 5
Train classifier to obtain accuracy of binary vector of n_i and save it in f_i
28 end
29 [*MaxFit*,*Index*]=max(*f*)
30 *B_Nest=*n_{index}
31 B_Fitness=MaxFit
32 while (*t* ≤ *Maxeval*) do
33 for *each nest* n_i (*i=1,. . . ,m*)
34 do
35 Calculate the Levy flights for generating new
solution with step size $\alpha \leftarrow A/\sqrt{G}$
store new solution into $nnew_i$
36 end
37 for *each nest* $nnew_i$ (*i=1,. . . ,m*)
38 do
39 for *each dimension j* (*i=1,. . . ,m*)
40 do
41 if $x_i^j > U_b$ then
42 $x_i^j = U_b$
43 else if $x_i^j < L_b$
44 else $x_i^j = x_i^j$
45 end
end
46 Convert $nnew_i$ to binary vector using Eq: 4&5
Train the classifier to obtain accuracy of binary vector of $nnew_i$, save it in *acc*
47 if $acc > f_i$ then
48 $f_i = acc$
49 $n_i = nnew_i$
50 end
51 end
52 [*MaxFit*,*Index*]=max(*f*)
53 *B_Nest=*n_{index}
54 *B_Fitness=MaxFit*
55 end

The N subset of features which are the output value of CMIM technique, the IBCS algorithm takes it as its input data.

As the problem of FS is considered as binary discrete problem, and each nest represent a solution, the value of nest is randomly selected and converted into binary vector as follows:

If the value of σ (as shown in eqn. 5) is equal to 1, then feature (SNP) is selected.

If the value of σ (as shown in eqn. 5) is equal to 0, then feature (SNP) is discarded.

The next improvement takes place in the step size scaling factor α of Levy flight equation as to increase the localized searching capability. In standard CS and BCS the value of scaling factor is constant, α=1. The new value is calculated as: $\alpha \leftarrow A/\sqrt{G}$ where A is step size and G stands for the generation number.

Complex Disease Datasets

The SNPs datasets represents the mutation in the sequence of nucleotide on a specific region for any group of individuals. The SNPs microarray data of five Affymetrix mapping 250 k arrays have been used in this study. The series of SNPs array are: GSE9222 [14], GSE13117 [15], GSE67047 [13], GSE34678 [10] and GSE16125 [17], all are downloaded from NCBI GEO (Gene expression omnibus) database [3].

The NCBI GEO is a free public repository that contains high throughput genomic and next generation sequencing data. Each dataset consist of two labels, Case and control to represent the affected individuals and healthy one. Every data sample represents its genotype at specific loci. The four SNP arrays, GSE9222, GSE13117, GSE67047 and GSE34678 are having data in alphabetical format except GSE16125 SNP, which has values in real numbers as shown in Table 1.

To apply classification and feature selection, the alphabetical dataset must be converted into numerical format.

There are numerous methods to convert into numeric, in this article the encoding used is: AB, AA, BB, and NO Call to 01, 11, 10 and 00 respectively. The results are compared with results achieved by [1] on same SNPs datasets.

Classification and Evaluation Measures

The SVM (support vector machine is used as a learner to calculate the performance measures of a proposed technique. The SVM is a supervised learning technique proposed by Boser et al. [4].

A k-fold cross validation is used where the value of k is set to k=2, where cases are less than 500 and k=4, where the cases are above 500 in the selected datasets. With the use of k-fold cross validation the proposed method have enough amount of cases for testing and training process.

The performance of each fold measure by following factors:

Table 1. Summary of SNPs datasets.

Dataset (GEO Series No.)	Information	No. of SNPs	No. of samples
GSE9222	ASD(Autism)	250000+	567
GSE13117	Mental Retardation	250000+	360
GSE67047	Thyroid cancer	1000000	225
GSE34678	Colorectal cancer	250000+	124
GSE16125	Colon cancer	250000+	42

Table 2. Parameters of IBCS algorithm

Parameters	λ	α	P_a	Pop. Size	Iterations
Values	1	$A/\sqrt{G}$	0.35	25	10

Table 3. No of SNPs before and after applying CMIM

Dataset	Total No SNPs	SNPs Selected by CMIM	No.of samples
GSE9222	250000+	500	567
GSE13117	250000+	500	360
GSE67047	1000000	1000	225
GSE34678	250000+	500	124
GSE16125	250000+	500	42

Table 4. Performance comparisons of ASD data.

GSE9222(ASD)	SNPs (Features)	Accuracy
ReliefF+SVM	60	0.782
CBFS+SVM	10	0.643
CMIM+SVM-RFE	100	0.895
RFS+SVM	10	0.642
Proposed(CMIM+IBCS)	**50**	**0.906**

Table 5. Performance comparisons of MR data

GSE13117(MR)	SNPs(Features)	Accuracy
ReliefF+SVM	30	0.781
CBFS+SVM	70	0.862
CMIM+SVM-RFE	50	0.850
RFS+SVM	10	0.731
Proposed(CMIM +IBCS)	**60**	**0.905**

Table 6. Performance comparisons of CR data

GSE34678(CR)	SNPs (Features)	Accuracy
ReliefF+SVM	40	0.675
CBFS+SVM	60	0.812
CMIM+SVM-RFE	50	0.903
RFS+SVM	20	0.712
Proposed(CMIM+ IBCS)	**50**	**0.926**

Table 7. Performance comparisons of CC data

GSE67047(Colon cancer)	SNPs (Features)	Accuracy
ReliefF+SVM	40	0.708
CBFS+SVM	30	0.850
CMIM+SVM-RFE	50	0.841
RFS+SVM	50	0.708
Proposed(CMIM+ IBCS)	**60**	**0.872**

Table 8. Performance comparisons of Thyroid cancer data

GSE16125(Thyroid)	SNPs (Features)	Accuracy
ReliefF+SVM	50	0.782
CBFS+SVM	60	0.754
CMIM+SVM-RFE	50	0.901
RFS+SVM	20	0.712
Proposed(CMIM+ IBCS)	**50**	**0.916**

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (6)$$

$$F = 2 \times \frac{Pre \times Re}{Pre+Re}, \quad (7)$$

where:

$$Re = \frac{TP}{TP+FN} \times 100, \quad (8)$$

$$Pre = \frac{TP}{TP+FP} \times 100, \quad (9)$$

where TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative.

We calculate the overall performance by using average of all folds:

$$Z = \frac{1}{K} \sum_{i=1}^{k} Accuracy\ i, \quad (10)$$

where *K* is total number of folds.

The proposed algorithm also uses an objective function for achieving maximum accuracy. Let us consider that N is total number of features, P is selected features by metaheuristic techniques, α is weight given to features and its value set to be $\alpha = 0.2$. Considering the above mentioned factors the value of fitness function $f(x)$ is:

$$f(x) = \left(\alpha * \frac{P}{N}\right) - \left((1-\alpha) * Z\right). \quad (11)$$

The proposed technique executes on Matlab 2018(a) on windows 10, Intel dual core i5, 2.5GHz and 8GB RAM.

3 Results and Discussion

The following tables provide the comparative results of proposed and other feature selection algorithm that are applied on above mentioned datasets.

The reduced feature subset of each SNPs dataset after applying fast CMIM filter technique are shown in Table 3.

The fast CMIM technique rank and arrange the feature subset in decreasing order of their feature score, we pick top 500 or 1000 feature according to the feature size. The comparison of different feature selection and proposed method on different disease SNPs datasets are shown in Tables 4, 5, 6, 7 and 8.

The results comparison from all the tables' shows that the proposed method outperforms all the other feature selection method in terms of accuracy. The performance comparison shown in

Table 9. Comparison of metaheuristic techniques with IBCS

Dataset	Objective Function	Algorithm	Accuracy
GSE9222(ASD)	*z*	BPSO	0.835
		BACO	0.845
		BGA	0.873
		IBCS	**0.862**
	F(x)	BPSO	0.852
		BACO	0.848
		BGA	0.871
		IBCS	**0.906**
GSE67047(Colon cancer)	*z*	BPSO	0.782
		BACO	0.793
		BGA	0.825
		IBCS	**0.863**
	F(x)	BPSO	0.795
		BACO	0.806
		BGA	0.830
		IBCS	**0.872**
GSE16125(Thyroid)	*z*	BPSO	0.819
		BACO	0.804
		BGA	0.836
		IBCS	**0.893**
	F(x)	BPSO	0.824
		BACO	0.816
		BGA	0.847
		IBCS	**0.916**

the tables are based on optimal number of SNP subset. As in dataset ASD, CR and Thyroid cancer the optimal feature subset is 50 SNPs, where in CC and MR datasets the feature subset is 60 SNPs. In some cases as in MR and colon cancer the accuracy of proposed method rise more than 3 to 4%.

After reducing the feature subset using fast CMIM, the IBCS algorithm was also compared with different metaheuristics techniques that are being used for feature selection.

With reference to classification performance the IBCS outperformed all other algorithms while considering average accuracy using the different fitness function, Z in equation 11 and $f(x)$ in equation 12 are shown in table 9. In GSE9222 (ASD) the BGA algorithm shows better performance with Z function.

4 Conclusion

The proposed algorithm in this article, CMIM and (IBCS) Improved binary cuckoo search optimization aims to classify complex diseases SNPs in the datasets.

The CMIM technique was used to reduce the large feature space of SNPs data, after that IBCS algorithm was applied to select best feature from feature subset and maximize the classification performance.

Our experiment was conducted on five different SNPs datasets of complex diseases obtained from NCBI GEO, and the proposed method shows higher accuracy then all the other compared methods like ReliefF, CBFS, RFS and CMIM+SVM-RFE.

After reducing the large feature space by CMIM technique the feature subset used by IBCS was compared with other metaheuristics method such as Binary Ant Colony Optimization, Binary Genetic Algorithm, and Binary Particle Swarm Optimization.

The results revels that IBCS achieved better accuracy as compared to other algorithms. The Support vector machine is used for performance measure of feature subsets; however other classifier can also be used.

The obtained result shows that the SNPs of the whole genome could be applied to find out the affected person from the healthy ones.

For future our proposed algorithm can be applied to other large biomedical datasets rather than SNPs data and also modified for the feature selection of the multiclass datasets. There is also a large amount of work required in order to understand the genetic basis of complex diseases.

References

1. **Alzubi, R., Ramzan, N., Alzoubi, H., & Amira, A. (2018).** A hybrid feature selection method for complex diseases SNPs. *IEEE Access*, Vol. 6, pp. 1292–1301. DOI:10.1109/ACCESS.2017.2778 268.
2. **Anekboon, K., Lursinsap, C., Phimoltares, S., Fucharoen, S., & Tongsima, S. (2014).** Extracting predictive SNPs in Crohn's disease using a vacillating genetic algorithm and a neural classifier in case–control association studies. *Computers in Biology and Medicine*, Vol. 44, pp.57–65. DOI:10.1016/j.compbiomed.2013.09. 017.
3. **Barrett, T. & Edgar, R. (2006).** Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods in Enzymology*, Vol. 411, pp. 352–369. DOI:10.1016/S0076-6879 (06)11019-8.
4. **Boser, B.E., Guyon, I.M., & Vapnik, V. N. (1992).** A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT'92*, pp.144–152. DOI:10.1145/130385.130401.
5. **Boutorh, A. & Guessoum, A. (2016).** Complex diseases SNP selection and classification by hybrid association rule mining and artificial neural network-based evolutionary algorithms. *Engineering Applications of Artificial Intelligence*, Vol. 51, pp. 58–70. DOI:10.1016/j.engappai. 2016. 01.004.
6. **Zheng, C.H., Huang, D.S., Zhang, L., & Kong, X.Z. (2009).** Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Transactions on Information Technology in Biomedicine*, Vol. 13, No. 4, pp. 599–607. DOI: 0.1109/TITB.2009. 2018115.
7. **Collins, F.S., Brooks, L.D., & Chakravarti, A. (1998).** A DNA Polymorphism discovery resource for research on human genetic variation: table 1. *Genome Research*, Vol. 8, No. 12, pp. 1229–1231. DOI:10.1101/GR.8.12.1229.
8. **Deutsch, J.M. (2003).** Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*, Vol. 19, No. 1, pp. 45–52. DOI:10.1093/bioinformatics/19.1.45.
9. **Fleuret. F. (2004).** Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, Vol. 5, pp. 1531– 1555.
10. **Jasmine, F., Rahaman, R., Dodsworth, C., Roy, S., Paul, R., Raza, M., Paul-Brutus, R., Kamal, M., Ahsan, H., & Kibriya, M.G. (2012).** A genome-wide study of cytogenetic changes in colorectal cancer using SNP microarrays: opportunities for future personalized treatment. *PLoS ONE*, Vol. 7, *No.* 2, pp. e31968. DOI:10.1371/journal.pone.003 1968.
11. **Khan, K. & Sahai, A. (2013).** Neural-based cuckoo search of employee health and safety (HS). *International Journal of Intelligent Systems and Applications*, Vol. 5, No. 2, pp. 76–83. DOI: 10.5815/ijisa.2013.02.09.
12. **Knudsen, S. (2011).** *A Biologist's Guide to Analysis of DNA Microarray Data*. John Wiley & Sons, Inc.
13. **Luzón-Toro, B., Bleda, M., Navarro, E., García-Alonso, L., Ruiz-Ferrer, M., Medina, I., Martín-Sánchez, M., Gonzalez, C.Y., Fernández, R.M., Torroglosa, A., Antiñolo, G., Dopazo, J., & Borrego, S. (2015).** Identification of epistatic interactions through genome-wide association studies in sporadic medullary and juvenile papillary thyroid carcinomas. *BMC Medical Genomics*, Vol. 8, No. 1, pp. 83.
14. **Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B.,**

Fiebig, A., Schreiber, S., Friedman, J., Ketelaars, C.E.J., Vos, Y.J., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., & Scherer, S.W. (2008). Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, Vol. 82, No. 2, pp. 477–488. DOI:10.1016/j.ajhg.2007.12.009.

15. **McMullan, D.J., Bonin, M., Hehir-Kwa, J.Y., de Vries, B.B.A., Dufke, A., Rattenberry, E., Steehouwer, M., Moruz, L., Pfundt, R., de Leeuw, N., Riess, A., Altug-Teber, Ã., Enders, H., Singer, S., Grasshoff, U., Walter, M., Walker, J.M., Lamb, C.V., Davison, E.V., & Veltman, J.A. (2009).** Molecular karyotyping of patients with unexplained mental retardation by SNP arrays: A multicenter study. *Human Mutation*, Vol. 30, No. 7, pp. 1082–1092. DOI:10.1002/humu.21015.
16. **Pereira, L.A.M., Rodrigues, D., Almeida, T.N.S., Ramos, C.C.O., Souza, A.N., Yang, X.S., & Papa, J.P. (2014).** *A Binary Cuckoo Search and Its Application for Feature Selection.* pp. 141–154, Springer.
17. **Reid, J.F., Gariboldi, M., Sokolova, V., Capobianco, P., Lampis, A., Perrone, F., Signoroni, S., Costa, A., Leo, E., Pilotti, S., & Pierotti, M.A. (2009).** Integrative approach for prioritizing cancer genes in sporadic colon cancer. *Genes. Chromosomes and Cancer*, Vol. 48, No. 11, pp. 953–962. DOI:10.1002/gcc.20697.
18. **Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., & Dante, M. (2001).** A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, Vol. 409, No. 6822, pp. 928–933.
19. **Salesi, S. & Cosma, G. (2017).** A novel extended binary cuckoo search algorithm for feature selection. *2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pp. 6–12. DOI:10.1109/ICKEA.2017. 8169893
20. **Shehab, M., Khader, A.T., & Al-Betar, M.A. (2017).** A survey on applications and variants of the cuckoo search algorithm. *Applied Soft Computing*, Vol. 61, pp. 1041–1059. DOI: 10.1016/j.asoc.2017.02.034
21. **Viswanathan, G., Bartumeus, F.V., Buldyrev, S., Catalan, J., Fulco, U., Havlin, S., da Luz, M.G., Lyra, M., Raposo, E., & Eugene-Stanley, H. (2002).** Lévy flight random searches in biological phenomena. *Physica A: Statistical Mechanics and Its Applications*, Vol. 314, No. 1–4, pp. 208–213. DOI:10.1016/S0378-4371(02) 01157-3.
22. **Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G.E., Raposo, E., & Stanley, H.E. (1999).** Optimizing the success of random searches. *Nature*, Vol. 401, No. 6756, pp. 911–914. DOI:10.1038/44831.
23. **Yang, X.S., & Deb, S. (2010).** Engineering optimization by cuckoo search. *International Journal Mathematical Modelling and Numerical Optimization*, Vol. 1, No. 4, pp. 330–343. DOI: 10.1504/IJMMNO.2010.035430.

Article received on 02/04/2020; accepted on 22/05/2020. Corresponding author is Manu Phogat.