Artificial Reproduction System and Imbalanced Dataset-A Mendelian Classification

Anita Kushwaha, Dr. R.S. Pandey

Abstract—We propose a new evolutionary computational model called Artificial Reproduction System which is based on the the complex process of meiotic reproduction occurring between male and female cells of the living organisms.Artificial Reproduction System is an attempt towards a new computational intelligence approach inspired by the theoretical reproduction mechanism,observed reproduction functions, principles and mechanisms.A reproductive organism is programmed by genes and can be viewed as an automaton,mapping and reducing so as to create copies of those genes in its off-springs.In Artificial Reproduction System,the binding mechanism between male and female cells is studied,parameters are chosen and a network is constructed also a feedback system for self regularization is established .The model then applies Mendel's law of inheritance,allele-allele associations and can be used to perform data analysis of imbalanaced data,multi-variate ,multiclass and big data.In the experimental study Artificial Reproduction System is compared with other state of the art classifiers like SVM,Radial Basis Function,neural networks,K-Nearest Neighbour for some benchmark datasets and comparison results indicate a good performance.

Index Terms— Big data, Bio-inspired Computation, Classification Algorithms, ,Homeostatis,Meiotic Reproduction ,Mendel's Law of Inheritance,Natural Computing

1 INTRODUCTION

Natural Computing [1,21] is an interdisciplinary field that formalizes complex processes occurring in living organisms to design various computational models for solving NP-hard problems or designing artificial systems with more natural behavior. It is inspired by biological course of action and is based on tools which are abstracted from natural phenomenon like brain-modeling, self evaluation, self-repetition, self-replication, immune system, Darwinian survival of fittest, granulation and perception. The computational model thus developed is based on various computing techniques like Artificial Neural Network, fuzzy logic, rough sets, evolutionary algorithms, fractal algorithms, DNA computing, granular computing or perception based computing. Biologically inspired computing is the sub domain of Natural inspired computing. According to De Castro and Von Zuben, the natural system and processes are further modeled and the natural system and processes are further modeled and simulated on computer and these theoretical models [9] provide in-depth insight and better understanding of the biological phenomenon and can develop computational system as well as algorithms to solve the complex problems. These techniques are also considered Heuristic or meta-heuristic[10,20] to the problems thatcannot be solved by other traditional techniques like linear, non-linear and dynamic programming.

There are a large number of bio-inspired computation approaches in the literature. Artificial Neural Networks (ANN) [11], evolutionary algorithms which are inspired by evolutionary biology [12, 13], artificial immune system which is inspired by the defense mechanisms of the antibodies to antigens [14] and immunology. Additionally growth and development processes of living organisms [15] are also considered as the bio-inspired computation.

The substantial achievements made in ANN and Artificial Immune system have shown an explicit importance in developing computational models of intelligent systems by utilizing the biological information processing mechanisms. These works have also motivated the research pertaining to other biological information processing systems. The Artificial reproduction model (ARS) which is developed forms the natural computing paradigm and is based on biological phenomenon of reproduction occurring in male and female reproductive cells.

Reproduction is the act of fertilization, crossover and reproducing off springs. Formally Reproduction is the act of 'reproducing' or creating multiple copies of stored programs or information to preserve the programs or information for future use. Reproduction may apply to the same organisms or between different organisms of the same species. Thus it can be asexual type or sexual type. Sexual type (like crossover, fertilization, zygote formation) occurring between two distinctly different

- *Anita Kushwaha is working as Assistant Professor at Birla Institute of Technolgy MESRA, Ranchi at its Allahabad Campus and is currently pursuing Doctorate degree in Computer Science From Dr.APJ Abdul Kalam Technical University Lucknow,India, PH-0532 2687363. E-mail: a.kushwaha@bitmesra.ac.in*
- *Dr. R.S.Pandey is currently working as Assistant Professor at Birla Institute of Technolgy MESRA,Ranchi at its Allahabad Campus PH-0532 2687363. E-mail: ravishankarbit@yahoo.com*

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

species like male and female species.

Predictive models of reproductive living cell are the most challenging task. Rosen [8] argued that life cannot be reduced to physical laws of universe as the reductionist models of life were not adequate for giving a formal description of the organization observed. Instead Rosen [8] suggested a high level description of life based on systems consisting of interacting components. This kind of approaches been adopted elsewhere in the artificial life community including work by Adams and Lipson [3]. In [5, 6] based on Gibson's theory of affordances the reproducer classification is described and applied to special types of computer program-virus program.

Problems with reproduction are that Self-replication takes different forms in different living systems [2].Identifying Reproduction [2] is a direct result of observation and may vary with the observer's biases. Many observed form of Reproduction include multiple parents, multiple off-springs, dissimilarities between parents and off-springs. Many formal definition of Reproduction do not take these factors into consideration including definitions given by Von Newman .In [4] the author describes entity based model applicable to artificial life and how entities can be combined together in Langton's Loop to produce higher level entities. Langton's loop provides a simple mechanism of self reproduction in cellular automata. One of the requirements of Von Newman's self-reproducing automaton [4] is it should be capable of universal computation and construction but a reproducer like a biological cell shows neither.

ARS overcomes these problems by using a formal approach to develop the computational model for reproduction. The system considers a number of parameters and tries to give a generalized model. The ARS Computational Model is derived from reproductive cell's complex behavior. The self regularization (Homeostasis based on hormones) mechanism between reproductive cells based on a feedback system is also modeled. The model can be used as a classifier for identifying new patterns in multivariate and voluminous amount of Big Data. It can successfully model the imbalanced data, multi-class data multi-variate data and Big Data. Another benefit is its efficiency and ease of use.

This research paper is organized as follows: the first section discusses the mathematical and the formal approach towards the construction of Artificial Reproduction System (ARS). The second section describes the actual system of Artificial Reproduction System Network and the classification rules with respect to four benchmark datasets. The third section describes the complete Artificial Reproduction System (ARS) which is a system of systems and also its applications to big data computing. We also present the experimental simulation and comparisons. Finally the fourth section gives the future scope and conclusion of the research work.

1.1 Formal Approach

The process of reproduction is binding together of the parent cells i.e male cell termed as M-cell with a female cell termed as F-cell in the reproductive environments (in –vivo or in-vitro) to produce a zygote or Z-cell which is a pre form of off spring. The process can occur in single M-cell and F-cell or in sets of multiple cells called M-set or F-set respectively, depending upon the naturally occurring reproductive rules (survival of the fittest) of living species. It can produce singles, doubles, quadruples or other multiple numbers of off springs in the next generation. M sets of M-cells and F-sets of F-cells that refer to multiple numbers of similar types of Male cells and Female cells respectively, participates in the reproduction process.

- A Reproduction model [2] specifies a state space and events that occur to move from one state to another.
- A basic reproduction model is a tuple

(S, A, →, Ent , r , ",P)

Where

(S, A, →) is a labeled transition system

Ent is a set of entities with r ∈ Ent the particular entity that reproduces in the model

" ⊆ Ent x S is a binary relation with e " s indicating that entity e is present in the state s;

- P is a path through the transition system representing the reproduction of r ,i.e,

P consists of sequence s0

a1→s1

a2→…

an → sn with si-1

ai → si for 0<i≤n

and with r " so and r " sn

The last item in this definition states that there is at least one path through r, e, and p

(s0, A0, ->) is a label led transition system

E is the set of entities

R ϵ E that reproduces with e ϵ S so entity E is present in S

P is a path through transition system representing reproduction of R

In asexual reproduction there is just one entity that can be identified as reproducer but in sexual reproduction [2] the genetic material that is transferred to the off-spring comes from two parents. To simplify we can consider that the entity that reproduces in sexual reproduction i.e.

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

a male and a female is rather a set {M, F} denoted by M+F. The reproducer will reproduce itself over time. Even if only a single male or a single female is produced during reproduction the requirement that reproducer is present in start and end state is fulfilled.

{M, F}→ M, F, M, F, M, F

{F, M}→ M, F, F, M, M, F

A complex reproductive unit R composed of an M cell and a F cell are present in the start state and they reproduce and generate a M or a F. In the end state a complete reproductive unit of M and F is still present. Reproductive units of three or more members can also be modeled in similar manner.

The reproduction process can be thought of as consisting of six steps or actions or events. The numbers of states that result are seven. Events or Actions are

1. AFFINITY (A):- coming closer of M-cell and F-cell based on affinity between receptors on surface of M-cell and F-cell and also distances between them.
2. BINDING (B):- Binding of M-cell with F-cell depending on the strength of their affinities.
3. CROSSOVER (C):- Crossover of genetic material contained in both parent cells.
4. SEGREGATION (S) :-Segregation or separation of alleles of off spring
5. ZYGOTE (Z) :- Formation of Zygote Z cell
6. OFFSPRING (O):- Conversion of Z-cell into off spring.

The states and events or action are as given under :

S1—A—S2—B—S3—C—S4—S—S5—Z—S6—O—S7

1.2 Shape Space

The theory of dynamical system contains a wide variety of tools for analyzing the non linear systems. We make use of shape space to describe quantitatively the bindings that occur between M-cell and F-cell of a reproductive unit R. The surface receptors present on both M-cell and F-cell are known to contain ducts for releasing binding chemicals to increase the affinity between two cells that determines the fertility rate. The binding strength also depends on distance between the two cells so the extent to which the two cells bind together depends on their shapes, the distance between them and the fertility rate. To model this matching we represent both M-cell and F-cell as a point in shape space of some dimension and affinity between them is measured by specifying a measure of distance between these two points. It is appropriate to consider discrete shapes space rather than continuous Floating point representation although it gives finer level of granularity but is course grained on larger scales. Hamming shape space metric or Euclidean shape space metric can be chosen to model binding strengths of M-cell and F-cell but each comes with its own bias so an appropriate measure ought to be chosen.

2 .The System of Systems

2.1 The Artificial Reproduction System (ARS)

The system is complex as it consists of two subnets of Female Reproductive cells Network (FRCN) and a Male Reproductive Cells Network (MRCN). Every FRCN as well as MRCN is associated with two feedback networks N1 and N2 for self-regularization based on hormones secretion.

Algorithm

1. Create a FRCN with n input F-cells.
2. Initialize it to F_i=(F1…Fm) random F-cells and M_i=(M1…Mn) random M-cells
3. Next, iterate through input attributes of training set and compute the output of the input layer
   $Y_i=F_i + M_i$.
4. This will be the input for the hidden layer of Zygote.
5. Create n hidden Zygote cells and supply $I_z=Y_i$.
6. Compute the output of hidden layer G_i ($G_i=I_z$).This will create reproductive changes and produce input for output layer. This is DNA.
7. Find output of the output layer. This is the allele present on the DNA.

(Note: - $Y_i=F_i + M_i$ the output of this input layer consists of adding or subtracting a small quantity to F_i or to M_i).

It should be noted that during allele formation [16, 17] the transformations we take are-the first gene is always considered as gene in 5_cap and last gene is always considered as poly (A) tail. The order of generation of different components of alleles (group of genes) can vary. First 5_cap then poly (A) tail then 5_UTR afterwards 3_UTR and if genes are still available then we create the triplet mRNA codon. These mRNA are very helpful in exposing the hidden patterns of datasets and thus decision rules can easily be formulated.

Zygote contains genome (RNA and DNA) and also chromosomes occurring in double (diplod) which becomes half (haploid) in the off-springs. If we take 0 and 1 to stand for recessive and dominant alleles then there are two homozygous dominant (DD) and recessive (rr) respectively. One heterozygous (Dr or rD) allele is present with 1 or 0 allele appearing in off springs if and only if one of the parents contained only 1 or only 0 allele and

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

possibility of randomly owning 01 or 10 if both the parents contained both dominant and recessive alleles (00 or 11).

In one example let the phenotype for wrinkled and Green characteristics (traits) of living organisms be AABB. Let AA stand for wrinkled. 'A' for dominant allele 'a' for recessive allele. Let BB stands for Green. 'B' for dominant allele 'b' for recessive allele. Then according to Mendel's law of Inheritance [26,27,28,29,30] for allele-allele association [26,27,28] the following heterozygous combinations of traits (alleles) occurs which can be categorized into 4 distinct different classes (here four are :Class X, Class Y, Class Z, Class W) .This is presented in the form of Penneth's lattice.

Table 1:
Mendel Law of Inheritance presented in the form of Penneth lattice.

	AB	Ab	aB	ab
AB	AABB	AABb	AaBB	AabB
Ab	AABb	AAbb Y	AaBb	Aabb Y
aB	AaBB	AaBb	aaBB Z	aaBb Z
ab	AaBb	Aabb Y	aaBb Z	aabb W

Here {AAbb, Aabb, Aabb} with 'A' dominant and 'b' recessive forms class Y.

{aaBB, aaBb, aaBb } with 'a' recessive and 'B' dominant forms class Z

4{aabb} with 'a' recessive and 'b' recessive forms class W.

Rest of the all combinations with 'A' dominant and 'B' dominant forms class X. There is 9:3:3:1 chance for occurrence of each of X, Y, Z, W classes according to Mendel's law of Inheritance. Here in this example, we have considered only two traits or alleles i.e. Green and wrinkled. We can further increase the number of traits (alleles) under consideration to 3, 4 , 5,…and so on and accordingly the Penneth Lattice expands in the multiple of 2^n .Number of different distinct classes results according to Mendel's Laws of inheritance.

Now we will take four datasets namely Iris dataset, Wine recognition dataset, Breast cancer Dataset and New – Thyroid dataset from UCI repository of machine learning databases and extract the decision rules according to the proposed method. Some of these are imbalanced datasets showing high imbalance ratio-IR.[23,24,25] like Iris0(with IR 2.0) Wisconsin (with IR 1.86) New thyroid (with IR 5.14). These rules produce almost 100% correct classification.

2.2. Iris flower dataset

This dataset consists of 50 samples from each of the three species of Iris flowers i.e., Iris setosa, Iris virginica, and Iris versicolor. Four features are measured from each sample; these are the sepal length, sepal width and petal length, petalwidth.

Let

x = Gene in poly (A) tail,

y = First gene of 5_ UTR,

z = Second gene of 5_ UTR,

w = Gene in 5_ Cap.

The knowledge discovery from the proposed method on this dataset displays the following premises and this will produce almost 100% correct classification.

1. $\{x \rightarrow 0\} \rightarrow$ setosa.
2. $\{\{x \rightarrow 1\} \wedge \{z \rightarrow 2\}\} \rightarrow$ setosa.
3. $\{\{x \rightarrow 1\} \wedge \{z \rightarrow \neg (2 \wedge 6)\}\} \rightarrow$ versicolor.
4. $\{\{x \rightarrow 1\} \wedge \{z \rightarrow 6\}\} \rightarrow$ virginica.
5. $\{\{w \rightarrow 6\} \wedge \{(y, z) \rightarrow (2, 4) \vee (3, 4)\} \wedge \{x \rightarrow 2\}\} \rightarrow$ versicolor.
6. $\{\{w \rightarrow 5\} \wedge \{(y, z) \rightarrow (3, 4)\} \wedge \{x \rightarrow 2\}\} \rightarrow$ versicolor.
7. $\{\{w \rightarrow 6\} \wedge \{(y, z) \rightarrow (3, 5)\} \wedge \{x \rightarrow 2\}\} \rightarrow$ (79% chance of virginica) $\wedge$ (21% chance of versicolor).
8. $\{\{x \rightarrow 6\} \wedge \{(y, z) \rightarrow (2, 5)\} \wedge \{x \rightarrow 2\}\} \rightarrow$ (75% chance of virginica) $\wedge$ (25% chance of versicolor).
9. $\{\{w \rightarrow 7\} \wedge \{(y, z) \rightarrow (3, 5)\} \wedge \{x \rightarrow 2\}\} \rightarrow$ (50% chance of virginica) $\wedge$ (50% chance of versicolor).

2.3. Wine dataset

There are three classes in this dataset and the number of instances in classes 1, 2 and 3 are 59, 71 and 48 respectively. The numbers of attributes are 13.

Let

x = First gene of RNA codon

y = First gene of 5_ UTR,

z = Gene in 5_ Cap,

p = First Gene in 3_ UTR,

q = Gene in poly (A) tail,

σ = RNA codon.

Given below are the decision rules for almost 100% correct classification.

1. $\{x \rightarrow (2 \vee 3 \vee 4)\} \rightarrow$ class-1.
2. $\{\{\sigma \rightarrow [2, 0, 1]\} \wedge \{y \rightarrow (1 \vee 4)\}\} \rightarrow$ class-2.
3. $\{\{\sigma \rightarrow [2, 0, 1]\} \wedge \{z \rightarrow 3\}\} \rightarrow$ class-2.
4. $\{\{\sigma \rightarrow [2, 0, 1]\} \wedge \{z \rightarrow (4 \vee 5)\}\} \rightarrow$ class-1.
5. $\{\{\sigma \rightarrow [3, 0, 2]\} \wedge \{y \rightarrow (1 \vee 6)\}\} \rightarrow$ class-2.
6. $\{\{\sigma \rightarrow [3, 0, 2]\} \wedge \{p \rightarrow 3\}\} \rightarrow$ class-2.
7. $\{\{\sigma \rightarrow [2, 0, 2]\} \wedge \{p \rightarrow (2 \vee 3)\}\} \rightarrow$ class-2.
8. $\{\{\sigma \rightarrow [2, 0, 3]\} \wedge \{p \rightarrow 3\}\} \rightarrow$ class-2.
9. $\{\{\sigma \rightarrow [2, 0, 2]\} \wedge \{z \rightarrow (3 \vee 4)\}\} \rightarrow$ class-2.
10. $\{\{\sigma \rightarrow [3, 0, 3]\} \wedge \{p \rightarrow 2\}\} \rightarrow$ class-2.
11. $\{\{\sigma \rightarrow [4, 0, 2]\} \wedge \{p \rightarrow 4\}\} \rightarrow$ class-2.
12. $\{\sigma \rightarrow [2, 1, 2] \vee [2, 1, 1] \vee [3, 0, 4] \vee [5, 0, 2]\} \rightarrow$ class-2.
13. $\{\sigma \rightarrow [1, 0, 0] \vee [1, 1, 0]\} \rightarrow$ class-2.
14. $\{\{\sigma \rightarrow [1, 0, 2]\} \wedge \{l \rightarrow (2 \vee 3)\}\} \rightarrow$ class-2.
15. $\{\{\sigma \rightarrow [1, 0, 1]\} \wedge \{l \rightarrow 3)\}\} \rightarrow$ class-2.

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

16. {{σ → [1, 1, 1]} ∧ {l → 6} ∧ {q→ 9}} → class-2.
17. {{σ → [1, 1, 2]} ∧ {l → (3 ∨ 4)}} → class-2.
18. {{σ → [1, 0, 1]} ∧ {l → 2} ∧ {q → 4}} → class-2.
19. {∀ Conditions→ ¬ (1 to 18)} → class 3.

2.4. Breast cancer dataset

This dataset has 32 attributes. Second attribute indicates the type of cancer so we ignored it out, and the remaining 31 are the real valued features. Total instances in this dataset are 569 of which 357 are of "Benign" and 212 are of "Malignant type.

Let
x = Last gene of second RNA codon
y = Last gene of seventh RNA codon
a = First gene of 3_ UTR
b = Last gene of 3_ UTR
p = 3rd RNA codon
q = 6th RNA codon
r = Gene in 5_ Cap
s = First gene of 5_ UTR
t = Second gene of 5_ UTR

Following premises produce 100% correct classification of breast cancer dataset.

1. {x → 1} → M.
2. {{x → 0}∧{(y → 1)∨((a∨b) → 1)}} → M.
3. {{x → 0} ∧ {(y → ¬ 1) ∨ ((a ∨ b) → ¬ 1)}}→ B.
4. {p → [2, 6, 4] ∨ [2, 3, 7] ∨ [0, 2, 8]} → M.
5. {{p → [1, 2, 5]} ∧ {q → [1, 7, 7] ∨ [2, 6, 3] ∨ [7, 9, 5] ∨ [2, 3, 9] ∨ [1, 8, 1]}} → M.
6. {{p → [1, 2, 4]} ∧ {q → [8, 1, 5] ∨ [9, 5, 8] ∨ [4, 7, 6]}} → M otherwise B.
7. {{p → [1, 3, 8]} ∧ {q → [2, 1, 5] ∨ [1, 5, 4] ∨ [6, 2, 8] ∨ [1, 3, 4]}} → M otherwise B.
8. {{p → [0, 1, 2]} ∧ {q → [9, 1, 6]}} → M otherwise B.
9. {{p → [1, 3, 9] ∨ [1, 2, 3] ∨ [1, 2, 8] ∨ [1, 1, 5] ∨ [1, 3, 5] ∨ [1, 3, 4] ∨ [2, 3, 3] ∨ [1, 2, 2] ∨ [2, 2, 9]∨[1, 3, 7]∨ [1, 3, 3]∨[1, 1, 2]∨[1, 2, 7]∨[0, 2, 2]∨ [2, 3, 7] ∨ [1, 4, 7] ∨ [1, 1, 1] ∨ [2, 4, 4]∨[1, 2, 9]∨[1, 3, 1]∨[1, 2, 1]∨ [1, 3, 6]∨[2, 3, 6]∨ [2, 3, 9] ∨ [1, 3, 2]} ∧ {q → [5, 1, 9] ∨ [8, 1, 4]∨[1, 4, 6]∨[4, 2, 6]∨[7, 3, 3]∨[9, 5, 1]∨[9, 6, 5]∨ [5, 4, 8] ∨ [4, 8, 6] ∨ [8, 6, 7] ∨ [2, 4, 2] ∨ [7, 3, 7]∨ [7, 3, 2]∨[1, 9, 1]∨[2, 8, 5]∨[3, 7, 3]∨[9, 4, 8]∨ [7, 2, 9] ∨ [7, 6, 7] ∨ [7, 8, 9] ∨ [8, 9, 5] ∨ [7, 4, 2]∨[6, 7, 5]∨ [8, 8, 1]∨[9, 3, 9]∨[1, 8, 8]∨[1, 7, 1]∨[5, 6, 2]∨[5, 8, 2]∨ [2, 7, 5]∨[6, 3, 9]∨[3, 8, 4]∨ [7, 3, 3] ∨ [7, 1, 2] ∨ [9, 3, 1] ∨ [8, 8, 7] ∨ [7, 4, 9] ∨[2, 4, 7] ∨ [3, 8, 9] ∨ [2, 6, 3] ∨ [9, 7, 1] ∨ [8, 4, 6]∨ [1, 4, 7]∨[7, 5, 3]∨[8, 1, 2]∨[3, 4, 5]∨ [9, 1, 6]}} → M otherwise B.
10. {{r → (7 ∨ 1)} ∧ {(s, t) → (9, 8) ∨ (2, 9) ∨ (4, 5) ∨ (4, 2)}} → B.

2.5. New-thyroid dataset

This dataset consists of three classes (1.Normal, 2. Hyper and 3. Hypo) with 215 instances in addition to six attributes. The

First attribute declares the belongingness of the class so we omit it out and thus the effective attributes are five. All attributes are continuous. Let

x = Gene in 5_ Cap
y = First gene of 5_ UTR
z = Second gene of 5_ UTR
a = Third gene of 5_ UTR
b = Gene in poly (A) tail

The following 10 premises given below are sufficient to produce 100% correct classification of this dataset.

1. {{(a, b) → [2, 0]} ∧ {(x, y, z) → [5,2, 3]∨[7, 5, 4]∨[8, 8, 4]}} → hyper otherwise normal.
2. {{(a, b) → [1, 0]} ∧ {(x, y, z) → [4, 1, 1] ∨ [4, 1, 2] ∨ [1, 1, 2] ∨ [1, 2, 2] ∨ [4, 1, 3] ∨ [4, 2, 2]}} → normal otherwise hyper.
3. {(a, b) → [1,−1]} → hyper.
4. {{(a, b) → [2, −1] ∨ [0, 0] ∨ [1, 1]} ∧ {(x, y, z) → [4, 5, 7] ∨ [7, 4, 3] ∨ [6, 8, 2] ∨ [9, 4, 3]}} → hyper otherwise normal.
5. {{z → 1} ∧ {(a, b) → [7, 1] ∨ [5, 6] ∨ [3, 3] ∨ [9, 6] ∨ [5, 2] ∨ [5, 5] ∨ [1, 5] ∨ [3, 2] ∨ [8, 8] ∨ [6, 2] ∨ [5, 1] ∨ [9, 8] ∨ [7, 5] ∨ [4, 6] ∨ [1, 5] ∨ [2, 4] ∨ [1, 7]}} → hypo.
6. {{a, b) → [2, 1] ∨ [2, 4] ∨ [2, 3]} ∧ {(x, y, z) → [9, 4, 1] ∨ [2, 1, 1] ∨ [1, 6, 1]}} → hypo otherwise normal.
7. {{z → 2} ∧ {(a, b) → [9, 5]∨[2, 4]∨[8, 4]}} → hypo.
8. {{(a, b) → [1, 3] ∨ [1, 4] ∨ [3, 4]} ∧ {(x, y, z) → [3, 7, 2] ∨ [3, 3, 2] ∨ [1, 4, 2]}} → hypo otherwise normal.
9. {{z → 0} ∧ {a, b) → [3, 9] ∨ [6, 8]∨[5, 4]}} → hypo otherwise normal.
10. {∀ Conditions→ ¬ (1 to 9} → normal.

3 The Complete system:

3.1 The Network of subnets:

The system is complex as it consists of two subnets a Female Reproductive cells Network (FRCN) and a Male Reproductive cells Network (MRCN).Every FRCN as well as MRCN is associated with 2 feedback networks N1 and N2 for self regularization. The FRCN as well as MRCN is connected with hormone secretion which is regulated by N1 and N2 feedback network. If hormone secretion increases then more number of male cells M-cells, indicated by opening the duct of N1, are increased. Similarly increased number of female cells F-cells , indicated by opening of duct of N2, are produced but if the hormone secretion decreases, less number of M-cells and F-cells will be produced as indicated by the closing of duct of N1 and N2.(Homeostasis).These 4 subnets forms the part of the complete ARS system.

A Genome (DNA or RNA) is designed to have two intertwined chromosomes stored as an array. In one chromosome, the attributes of each sub network as no. of units,

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

time constant and indices of other sub network to which this sub network projects is stored. The other chromosome contains learning rules which are used to adjust weights between individual units. The two chromosomes from parent are independently recombined. A hybrid chromosome with information from both the parents is created. A point is randomly chosen and information up to that point is copied from parental chromosome and rest of information comes from maternal chromosomes. A sub network and learning rule chromosome must be of same length for recombination to occur. A complex architecture can be represented by sub network connecting matrix. A positive number indicated sub network are fully connected and integer specifies which one of the many learning rules to be used for that connection.
These alleles or chromosomes are useful in exploring hidden pattern in dataset and decision rules can be constructed for different datasets.

3.2 Big Data Computing

Each reproducer (M-cell or F-cell) can be represented by a binary string, each bit corresponding to trait or a genetic marker representing loci which encoded various traits. Let's say there are six traits ABCDEF. Loci were divided into two chromosomes each with these three traits. Loci are mapped to phenotype which determines the behavior of reproducers.

M-cell A B C D | E F
F- Cell G H I J | K L crossover
Z-cell A B C D | K L
G H I J | E F

A network model of single ARS and a multi - ARS is discussed. A feedback system based on self regularization or homeostasis is also presented.
The following assumptions are taken for the model.
1. The reproducer DNA consists of 23 chromosomes. (K)
2. The no of loci may vary. (I)
3. The initial frequency of desirable allele (L)
4. The mutation rate (M)
5. The intensity of selection (s)
6 There is (genome, allele) pair that is mapped in each reproducer similar to (key, value) pair where each of the unique genome either DNA or RNA is the key or signature of the individual and Dominant or recessive value is the allele.
7. Reshuffling of genes is automatically done in crossover.

The basic states of reproduction model remain same. Only change is how map-reduce programming paradigm is implemented.

The Map Reduce consists of the two Phases.
1) Map and 2) Reduce.

The Map is used to split the job into several independent units and each unit is assigned to different computing data node. In the reduce phase, the data is aggregated, summarized, filtered or combined with the given data. The result is stored in a Distributed File System.
A reproductive unit R maps and reduce to a number of units called off-springs O in parallel. Offspring unit O can be single, doubles, quadruples or higher numbers depending upon number of nodes. A crossover point is selected. During crossover, the genome data is split into independent units. There are parallel map tasks, a shuffle or a sort phase and a parallel reduce task that runs after shuffling completes .A map function of input units takes place that output a set of intermediate records in the form (genome, allele) pair similar to (key ,value)pair. The map produces these output pairs, there is a split function partitions these pairs into R disjoint pairs by applying the function to genome alone or to the key only. There is also a reduce process which combines the pair assigned to it in some way and produces the final output. The process is highly parallelizable as a very large no R units can work in parallel. This map-reduces between reproductive units and offspring units can be easily implemented by writing map function and reduces function for various types of networks.
Mapper:
In parallel, Read records (genome,allele) pair same as (key, value) pairs from input files.
Apply filter or transformations to each instance
Call split function to partition records into R disjoint buckets
Write each bucket to processing node's local disk
Reducer:
Process or combine in parallel the records assigned to it based on same hash instance.
For the classification of data, we take two phases. There is a training phase that stores data vector coordinates along with class labels. The class label is used as identifier for data vector which can be used for classification in testing phase. In testing phase, the aim is to find class label for new points.

3.3 Simulation and Comparison

Now we perform different experiments in this section to show that this proposed algorithm is feasible and valid. The goal of our experiments are to evaluate and compare the test accuracies and running times by hard-margin accuracies and running times by hard-margins SVM, RBFNN, NN, and K-nearest neighbors(K-NN) with our proposed algorithm.
We have selected four databases from UCI repository of machine learning databases to perform the experiments. Details of these databases are shown in Table 2.

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

Table 2:
Characteristics of Dataset

Database Name	Number of instances	Number of Attributes	Number of classes
Iris	150	4	3
Wine	178	13	3
Breast Cancer	569	31	2
New Thyroid	215	5	3

We have taken the help of MATLAB for these four selected databases. Further, we have selected Gaussian hard-margin SVM for all datasets.SVM is a linear machine in which support vector Learning algorithm is used to implement the learning process using a given set of training data and it automatically determines the hidden units. RBFNN is a layered feed forward network with one hidden layer of RBF units and a linear output layer. All attributes are used as inputs and no weights are applied between input layer and the hidden layer i.e., default weights are 1's.The activation of a hidden unit is determined by the distance between the input vector and the center vector of the hidden units. The number of output neurons is taken equal to the number of class labels. The hidden neurons are taken equal to the twice of the input neurons. The NN we have taken is a three layer perceptron with gradient descent back propagation and number of hidden neurons is taken equal to two. The activation function for hidden layer is sigmoid. The learning rate and momentum is taken as 0.1 and 0.5, respectively. There is only one neuron in the output layer. We used the K-NN with Euclidean distance with *K* value 1.These results are evaluated by means of five-fold cross validation. In this five-fold cross validation, the datasets are divided into five disjoint subsets. Each time, we select one of the five subsets as the testing set and remaining four subsets are used as training sets. Then the average errors across all five trials are computed. From Tables 2 and 3, we see that these five classifiers require approximately the same training time. However, our proposed idea requires slight more training time than the hard margin SVM, RBF and NN but it is less than the training time of K-NN for Blood transfusion dataset. This proposed algorithm requires same training time as NN for Iris dataset but for New-Thyroid dataset, its training time is less than the K-NN. As a whole, the average testing accuracies of these classifiers over four datasets are 94.87% (hard-margin SVM), 95.14% (RBF), 91.22% (NN), 92.17% (K-NN) and 95.87% (ARS) respectively. ARS performs better than other classifiers for Iris and Wine datasets. Moreover, for wine dataset its performance is same as RBF neural network.

4. Conclusion and Scope of Future work

In this paper we have discussed the challenges and the problems in developing a new evolutionary computational model Artificial Reproduction System (ARS) based on the Reproduction process occurring in living cells and have experimentally shown that this computational model -ARS can be applied for classifying data that is imbalanced ,multivariate and multi-class and also voluminous i.e. Big Data. This computational model ARS paves a way for new developments in this field which is yet untouched and gives a new paradigm for computational model similar to artificial immune systems, artificial endocrine systems. It can also help to enhance the knowledge discovery and the model interpretation. The algorithm developed can be applied to various types of data like interval-valued data, multi-class, multi-variate, multi-modal, time-series and big data. It finds its application particularly in gene-expression data and medical data classification where volume of data is enormous and NP hard and data cannot be easily classified using standard classification algorithms. Additionally, we hope that our insights on many problems and challenges present in this relatively new research area will help guide the potential research directions for the future developments in thisfield.

Table 3:
Comparison between SVM , RBF and NN

		Guassian Kernal based hard SVM			RBF		NN	
Database Name	T(s)	Training (%)	Testing (%)	T(s)	Training (%)	Testing (%)	T(S)	Training (%)	Testing (%)
Iris	0.04	100	93.67	0.08	100	95.3	0.14	97.41	88.03
Wine	0.08	98.15	93.15	1.8	100	98.87	1.86	95.72	87.12
Breast Cancer	0.34	100	96.35	0.6	99.02	93.21	0.6	100	95.42
New Thyroid	0.06	100	96.34	0.04	98.03	93.21	0.6	98.51	94.33
Average	0.13	99.53	94.87	0.63	99.26	95.14	0.8	97.91	91.22

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

Table 4:
Classification data for K-NN and Artificial Reproduction System (ARS)

DataBase Name	T(s)	Training (%)	Testing (%)	T(s)	Training (%)	Testing (%)
Iris	0.04	94.32	91.52	0.14	100	97.36
Wine	2.26	98.71	93.11	2.64	100	98.34
Breast Cancer	0.58	96.04	90.32	0.64	97.62	94.90
New Thyroid	1.26	97.17	93.73	1.08	100	93.01
Average	1.03	96.56	92.17	0.95	99.40	95.87

Table 5:
FIVE-Fold cross validation of Artificial Reproduction System(ARS) for Iris Dataset

	1	2	3	4	5	Average
Training Pattern	120	120	120	120	120	120
Testing Pattern	30	30	30	30	30	30
Misclassification(testing)	0	1	1	1	1	0.8
Recognition rate(testing)	100	96.7	96.7	96.7	96.7	97.36

Table 6:
FIVE-Fold cross validation of Artificial Reproduction System (ARS) for Wine Dataset

	1	2	3	4	5	Average
Training Pattern	136	144	144	144	144	142.4
Testing Pattern	42	34	34	34	34	35.6
Misclassification(testing)	0	0	1	0	2	0.6
Recognition rate(testing)	100	100	97.10	100	94.10	98.24

Table 7:
FIVE-Fold cross validation of Artificial Reproduction System (ARS) for New Thyroid dataset

	1	2	3	4	5	Average
Training Pattern	172	172	172	172	172	172
Testing Pattern	43	43	43	43	43	43
Misclassification(testing)	3	2	4	3	3	3
Recognition rate(testing)	93.02	95.34	90.69	93.02	93.02	94.90

Table 8:
Comparison of 5 classification Algorithms SVM,RBF,NN, K-NN and Artificial Reproduction System(ARS)

	1	2	3	4	5	Average
Training Pattern	455	455	455	455	455	455
Testing Pattern	114	114	114	114	114	114
Misclassification(testing)	4	5	6	6	8	5.8
Recognition rate(testing)	96.49	95.61	94.73	94.73	92.98	94.90

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

Table 9:
Comparison of 5 classification Algorithms SVM,RBF,NN,K-NN and Artificial Reproduction System(ARS)

Dataset Class	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)	fp(%)
Iris	Setosa	4.31	96.32	3.44	95.131	3.67	91.14	6.61	92.61	2.34	97.61
	Versicolor	3.86	91.12	1.98	92.41	3.22	86.79	4.11	85.32	1.42	96.23
	Virginica	2.01	90.47	2.98	94.67	5.13	92.72	4.79	87.61	2.12	96.23
Wine	Class 1	3.11	92.01	4.73	94.01	6.17	91.84	6.11	91.05	1.72	96.64
	Class2	5.01	94.20	3.45	93.58	5.94	90.71	6.27	92.01	3.14	97.33
	Class 3	5.01	94.20	3.45	93.58	5.94	90.71	6.27	92.01	3.14	97.33
Breast Cancer	Benign	6.52	95.46	8.94	97.14	7.97	94.57	5.13	95.05	1.41	97.32
	Malignant	4.87	91.98	6.76	96.78	5.86	89.32	4.97	92.13	2.14	96.79
New Thyroid	Class1	5.03	94.21	3.88	88.49	3.89	84.78	4.97	92.13	2.14	96.79
	Class2	5.51	94.53	4.17	90.94	6.45	88.93	4.83	88.98	2.32	97.42
	Class3	4.37	93.23	4.57	91.68	7.21	92.13	5.12	93.03	1.89	96.66

Table 10:
Area under the Curve Classifiers

Dataset	Gaussian SVM	RBF	NN	K-NN	ARS
Iris	0.9694	0.9650	0.9456	0.9392	0.9806
Wine	0.9543	0.9580	0.9298	0.9328	0.9776
W Breast Cancer	0.9514	0.9514	0.9409	0.9497	0.9811

References

[1] G. Rozenberg, T. Back, J.N.KoK (Eds),Handbook of Natural Computing,Springer 2012.

[2] Matthew PaulWebster,"Formal Models of Reproduction: from computer viruses to Artificial Life",Ph.D Thesis submitted to University of Liverpool,2008.

[3] Adams.B,Lipson H,"Auniversal framework for self-replication," In Eurpean Conference on Artificial Life(ECAL'03)pages 1-9,2003.

[4] Matt Webster1,Grant Malcolm1,"Hierarchical Components and Entity based modeling in Artificial Life",Artificial Life XI 2008 678

[5] Amie Judith, Radenbaugh,"Applications of Genetic Algorithms in Bioinformatics",PhD theisi San Jose University,2008

[6] Matt Webster ,Grant Malcolm,"Formal affordanc based models for computerVirusReproduction", http://link.springer.com/article/10.1007%2Fs11416-007-0079-4

[7] Matt Webster ,Grant Malcolm,"Reproducer classification using Theory of affordances".In proceedings of 2007 IEEE Symposium on Artificial Life (CI-ALife 2007) pages 115-122.IEEE Press,2007.

[8] Rosen,R,"Essays on Life Itself",Columbia University Press. ISBN :978-0231105118

[9] Decastro L.N,Von Zuben F.J,"Recent development in biologically inspired computing",Idea Group Inc,pp 340-365.2005.

[10] Glover F,Kochenberger G,"Handbook of Metaheuristic",Kluwer Boston MA.2003.

[11] Haykins S,"Neural Networks:A Comprehensive Foundation ",2nd Edn,Pearson Education Asia.1999.

[12] Back T,Fogel D.B,Michalewicz,"Evolutionary Computation2 Advanced Algorithms and operation",Institute of Physics Publishing, Bristol (IOP)2000.

[13] Beyer H.G ,"Theory of Evolution Strategies",Springer-Verlag 2001.

[14] De Castro L.N,Timmis J."Arificial Immune Systems:A new Computational Intelligence Approach",Springer-Verlag Berlin.2002

[15] Kumar S,Bentlet P.J,"On Growth,form and Computers",Academic Press 2003.

[16] Doolittle R.F,"Molecular evolution:Computer analysis of protein and nucleic acid sequences" Meth. Enzymol 183.1990.

International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016
ISSN 2229-5518

[17] Snyder E.E and stormo G.D,"Identification of coding regions in genomics DNA",J.Comput. Biol 248,1-18.1995

[18] S.Bandyopadhyay,U.Maulik and D.Roy,"Gene Identification:classical and computational intelligence approaches",IEEE Transactions on Systems,Man,Cybernetics,Part C,Applications and Review,Vol 38,No.1,Jan 2008.

[19] E.E Snyder,G.D. stormo,"Identification of coding regions in genomic DNA sequences:an application of dynamic programming and neural networks",Nucleic acid Research Vol 21,no.3,pp.607-613.1993

[20] Jiming Liu,Kwok Ching Tsui,"Towards nature-inspired computing", Communications of the ACM 49(10):59-64 · October 2006

[21] Dana.H.Ballard,"An introduction to Natural computation", mitpress.mit.edu/books/

[22] Habibo He,member IEEE,Edwardo A Garcia,'Learning form imbalanced data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009

[23] John Nicolas Korecki,"Semi supervised Self learning on Imbalanced Data", A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science Department of Computer Science and Engineering College of Engineering University of South Florida,2010

[24] Rukshan Batuwita and Vasile Palade," Class Imbalance Learning Methods for Support Vector Machines", Singapore-MIT Alliance for Research and Technology Centre; University of Oxford

[25] V. López, A. Fernandez, S. Garcia, V. Palade and F. Herrera," An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics", Information Sciences 250 (2013) 113-141doi: 10.1016/j.ins.2013.07.007.

[26] G. Mendel, "Experiments on Plant Hybrids." In: The Origin of Genetics: A Mendel Source Book, (1866).

[27] C. Darwin, "On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life," p. 162 (1859)

[28] G. Mendel, "Mendel's Principles of Heredity." P.40 (1866).

[29] Gurmukh Singh, Khalid Siddiqui, Mankiran Singh and Satpal Singh, " Modeling Mendel's Laws on Inheritance in Computational Biology and Medical Sciences", Journal of Educational Technology Systems, Vol. 39, No. 1, 2010.

[30] Gurmukh Singh ,"Computer Simulations to Model Mendel's Laws on Inheritance in Computational Biology", ACISNR'10, May 5–7, 2010, Fredonia, New York, USA.