GENE SELECTION USING MULTIPLE QUEEN COLONIES IN LARGE SCALE MACHINE LEARNING

A. Sampath Kumar [a], P. Vivekanandan [b]

[a] Department of CSE, Sri Krishna College of Technology, India
[b] Department of CSE, Park College of Engineering and Technology, India.
Sampathkmr1987@gmail.com[a] and anandpvivek@yahoo.com[b]

Abstract— *In the field of bioinformatics research, there has been a tremendous increase in the volume of data. This is due to the fact that all the processes are digitized and there is an availability of high throughput devices at a lower cost owing to which data volume is rising everywhere. As an example, the approximate size of a single sequenced human genome is 200 gigabytes. With the growth of big data technologies, this trend in the increasing volumes of data is bolstered by reduced computing expenses and enhanced analytics throughput. Technologies such as automated genome sequencers that capture big data are becoming lesser expensive with increased efficacy giving rise to this new era of big data in the field of bioinformatics. There has been a supply of large volume of data in many fields due to the development of microarray technology. This has been especially useful in predicting as well as in the diagnosis of cancer. Since the extracted genes from microarray are rife with noise, the task is selecting genes that are related to cancer, so that the disease can be classified precisely. For the efficient feature selection in the Hadoop framework, a new feature selection algorithm has been suggested- Correlation based Feature Selection (CFS), Genetic Algorithm (GA) and Honey Bee Mating Optimization (HBMO) algorithm. These techniques help in decreasing the problem dimension and noise and improvising the algorithm speed by the removal of irrelevant or superfluous features. It has been shown by experimental outcomes that the suggested technique helps to achieve better performance when compared to the other techniques.*

Keywords*: Big Data, Bioinformatics, Microarray, Gene Selection, Feature Selection, Correlation based Feature Selection (CFS), Genetic Algorithm (GA) and Honey Bee Mating Optimization (HBMO).*

1 INTRODUCTION

One of the leading causes of death all around the world that accounts for more than 8 million deaths as per World Health Organization (WHO) is cancer. It is construed that these will rise to about 14 million in the next couple of decades. There are about a hundred know various cancer types; likely that the numbers are more. Cancer leads to a anomalous growth of cells leading to the development of a tissue called as mass which in turn attacks other organs in the body. Every year, breast cancer affects more than 1.3 million women throughout the world which also accounts for an average of about 14% of the deaths that are related to cancer.

1. There has been a dramatic rise in the increase in the incidences of breast cancer over the past decades; it is expected that these numbers will only increase in the coming years.
2. This is why breast cancer is expected to remain a considerable onus on healthcare [1].

Bioinformatics and Machine Learning: This involves data that collate gene related information from tissue and the cell samples that can aid the disease and specific tumor diagnosis. Even through extremely small samples exist for the purposes of

training as well as testing, which is about less than a hundred patients, the number of features may range from 6000 to 60,000 as it accounts for gene expression across the population. The binary approach is typical which separates the normal patients from cancer patients on the basis of their gene expression "profile". Data sets are also present and the aim here is to differentiate different types of tumors ; this is rather a more complex task. This is also referred to as the multiclass approach. Hence, for machine learning researchers, microarray data poses a challenge ; Since there are several fields corresponding to a very less number of samples, there is an increased likelihood of finding more "false positives" because of the chance- this impacts not only finding relevant genes but also in the construction of predictive models. For validating the models it is imperative to find superior methods and also to assess their probability. Additionally, complications involved in experiments like noise as well as variability lead to the domain of microarray data analysis as being exciting and also challenging [2].

There is a lot of genetic data that is comprised in the microarray databases; this improvises the understanding of medicine and biology when properly analyzed. There are several microarray experiments that have been designed for the investigation of genetic approaches of cancer; For distinguishing between cancerous and non cancerous and also for classifying various types of cancer, there have been several analytical approaches that have been applied. In microarray data analysis, several machine learning techniques have been looked into in the last decade. There have been several approaches that have been tried in order to – a) differentiate between normal and cancerous cells, (b) categorize different types of cancer and (3) Detect and identify the cancer subtypes that may aggressively progress. Thus, a biologically meaningful interpretation of complex datasets can be generated using these investigations that can drive the succeeding experimentation [3].

Selecting a subset of variables from the input that can describe the input data effectively by decreasing the noise or irrelevant data, yet provide good prediction outcomes is the objective of feature selection. One application of feature selection is in the analysis of gene microarray. There may be several hundred variables that may be correlated with other variables in a standardized gene expression data. An example is when two of the features are perfectly correlated, only one is sufficient for describing the data. As no extra information regarding the classes is provided by the dependant variables, it is a noise for the predictor. This means that a few of the unique features that comprise maximum differentiating information about the classes can provide the total information content. This removal of dependent variables causes a reduction in the amount of data that leads to an improvisation of classification performance. There are some applications wherein uncorrelated variables introduce noise thereby inducing bias in the predictor leading to a degradation of classification performance.

This takes place when there is a paucity of information regarding the process that is under investigation. Thus, some insight can be gained into the process by the application of feature selection techniques and this in turn improvises the requirement for computation as well as the accuracy of prediction [4].

The concept of big data, albeit extremely important does not conform to the conventional database structure. This is because data that is derived from the machine has a rich and diverse content that needs to be discovered and this data also multiples rapidly. Another example is the data that is obtained from social media which has rich textual content but is rife with meaningful insights. The challenges that are posed by vast, unstructured and fast moving data, that can be cumbersome for traditional data management, can be effectively handled by Big data analytics. Data which has

unprecedented scope and complexity is being generated from businesses and research institutions to government organizations. It has become globally important for organizations to be able to extract meaningful information for competitive advantages from huge amounts of data. It is extremely challenging to extract meaningful insights quickly and easily; This impacts both their business performance as well as their market share.In the recent years several tools have been made dispensable for handling the huge volumes, variety and the velocity of the data. These technologies are , however, not very expensive and mostly rely on open source software of which Hadoop is the most commonly employed framework that combines commodity hardware along with open source software [5].

Hadoop employs high level data processing languages. For handling petabytes of data across thousands of computers, Hadoop modules provide ease of language, graphical interface as well as administration tools. In today's world of big data processing, Hadoop and Map reduce are the two most commonly used models for processing big data. Hadoop makes use of simple programming models and is an open source large scale data processing framework .the framework supports the distributed processing of huge amounts of data. In addition to the other modules, the apache Hadoop project comprises the Hadoop map reduce as well as the Hadoop Distributed File System. While managing failures at the node level, the software can be modeled to harvest the processing power of clustered computing [6].

With the dawn of an era of big data which is complex and huge volumes of data , a critical role is played by feature selection that decreases the high data dimensionality in machine learning problems. Feature selection has the ability to describe a given problem with precision, without affecting the performance. It may seem theoretically attractive to have a huge number of input variables; but this in fact faces the issue of dimensionality that is internal not only to the data but is an issue associated with the data as well as the algorithm that is being used. This lead to the selection of features by the researchers in the pre-processing stage which can convert data into a lower dimension. These techniques for feature selection have developed of late and these are based on the relationship between feature selection algorithm as well as the inductive learning technique used for model inference-filter, wrapper and embedded are a few methods that are used. These techniques may also be categorized based on the individual computation and subset computational techniques. When evaluating the individual, it is referred to as feature ranking. This helps in evaluating the features of the individuals by the allocation of weights as per relevance. The subset evaluation on the other hand generates candidate feature subsets that are based on particular search strategy which may be evaluated by some or the other measure [7].

In this work, proposes the CFS, GA and HBMO algorithm based feature selection in bioinformatics and big data. The remainder of the work is structured thus: in section 2, related works in literature are discussed. In section 3, the materials and methods used in the presented work is explained. Section 4, discusses about the results, concluded in section 5.

2 RELATED WORKS

The main tenets of feature selection and the way in which they are being applied in the field of big data bioinformatics was discussed by Wang et al., [8].This technique pre-empted the use of filter, wrapper and embedded approached for feature selection; Instead, the feature selection technique was looked upon as a combinatorial optimization or search problem; The feature selection techniques have been classified into exhaustive search, heuristic search and hybrid techniques. The heuristic search techniques have been further categorized as those with or sans the data distilled feature ranking measures.

A novel framework for effectively analysing high dimensional economic big data was formulated

by Zhao et al., [9]. This framework has been based on novel distributed feature selection. This framework particularly combined the techniques of economic feature selection as well as econometric model construction so that hidden patterns are shown for economic development. There are three pillars on which functionality rests upon-(i) Preparing high quality economic data using new techniques in pre processing. (ii) for locating important and representative economic indicators from multidimensional data sets, a new distributed feature identification solution , and (iii) Capturing the hidden patterns for economic development using novel econometric models . It was shown by empirical outcomes by means of economic data collated in Dalian , China that this framework has shown excellent performance while analysing huge amounts of economic data.

Kong et al., [10] proposed the Jointly Sparse Discriminant Analysis (JSDA) to explore the key factors in breast cancer and extract the key features for improving the accuracy in diagnosis and prediction. JSDA introduces the jointly sparse regular term (i.e. L2, 1 norms term) to the criterion. A convergent iterative algorithm is designed to solve the optimization problem. It is shown that the proposed JSDA algorithm not only can learn the jointly sparse discriminant vectors to explore the key factors of the breast cancer in cancer pathologic diagnosis, but also can improve the diagnosis accuracy compared with the classical feature extraction and discriminant analysis algorithm. Experimental results on breast cancer datasets indicate that JSDA outperforms some well-known subspace learning algorithms in prediction accuracy, not matter they are non-sparse or sparse, particularly in the cases of small sample sizes.

Wan & Freitas [11] evaluated four hierarchical feature selection methods, i.e., Hierarchical Information Preserving Features (HIP), Multi-Resolution (MR), Simple Hierarchical Selection (SHSEL) and Global Terrorism Database (GTD), used together with four types of lazy learning-based classifiers, i.e., naïve bayes, tree augmented naïve bayes, Bayesian network augmented naïve bayes and K-Nearest Neighbors (KNN) classifiers. These popular hierarchical feature selection techniques have been compared not only with each other but also with a popular “flat” feature selection technique referred to as CFS. The dataset of this adopted bioinformatics comprises genes that are related to aging which can be employed as instances. It also comprises gene ontology terms that are employed as hierarchical features.It has been shown by experimental outcomes that the chosen HIP technique performs excellently well as far as the predictive accuracy and the robustness involved in instance classes having sufficiently imbalanced distribution are concerned.

For balancing the precision and the stability of feature ranking as well as prediction, a Max-Relevance-Max-Distance (MRMD) feature ranking method was proposed by Zou et al., [12]. The authors have tested the technique on two of the data sets for proving the efficacy on big data .The first is the benchmark data set that has high dimensionality referred to as image classification. The second that is an outcome of private research and has many instances is the protein-protein interaction prediction data. It was experimentally proven that this technique maintained the precision along with the stability on both the huge data sets. Also compared to the other filtering and wrapping techniques like Minimum Redundancy Maximum Relevance (MRMR) and Information Gain (IG), this technique is much faster.

To select informative genes from microarray data sets a Maximum–Minimum Correntropy Criterion (MMCC) approach was proposed by Mohammadi et al., [13]. This approach was found to be stable, fast and also extremely resilient to diverse noise and outliers; This also gave better accuracy compared to the other algorithms. Additionally, an evolutionary optimization process was used for determining an optimal number of features contained in each data set. For about 25 commonly used microarray data sets, MMCC proved to be more efficient than the other popular

gene selection algorithms , and this was confirmed by broad experimental evaluation. Another surprising outcome was that the Support vector Machine (SVM) presented a better accuracy in classification by lesser than ten genes that the MMCC had chosen in all of the cases.

An innovative evolutionary technique that was based on the genetic algorithms and artificial intelligence for identifying predictive genes for the classification of cancer was suggested by Dashtban&Balafar [14].First, the dimensionality of the feature space was reduced by applying the filter method. This was followed by the incorporation of an integer coded GA having a couple of dynamic length genotype, intelligent parameter settings and altered operators; laplacian and fisher score which are the two popular filter techniques have been used , taking into account the following- similarities, quality of the genes that have been chosen and the influence that they have on the evolutionary technique. There were many statistical tests that accounted the selection of classifier, dataset and the filter method; Some considerable differences between the performance of various classifiers and the filter techniques over the data sets were exposed.

For teh sake of speeding up convergence, an innovative Gene Recombination Operator (GRO) was incorporated into the Artificial Bee Colony (ABC) algorithm by Li et al., [15]. Specifically speaking, in order to produce candidate solutions by gene combinations, in GRO, some part of the optimal solution from the current population has been chosen. This is true as every good solution joins with only one of the other good solution in order to generate a candidate solution. Additionally, only at the end of every generation, GRO will be initiated. The GRO has been incorporated in nine versions of the ABC for validating both the efficacy and the efficiency;That GRO could increase the usage of the ABCs and also speed up the convergence without compromising on diversity has been proven by experimental outcomes on 22 benchmark functions.

Sheikhpour et al., [16] proposed the Particle Swarm Optimization (PSO) and non-parametric Kernel Density Estimation (KDE) (PSO-KDE) based classifier to diagnosis of breast cancer. The PSO can find the bandwidth of the kernel and also choose the feature subset in the KDE based classifier at the same time, according to this suggested model. The criteria for designing the objective function of PSO-KDE are both the performance of classification and the number of features that have been selected. Using classification accuracy, sensitivity and specificity, the performance of PSO-KDE has been studied on Wisconsin Breast Cancer Dataset (WBCD) and Wisconsin Diagnosis Breast Cancer Database (WDBC). It has been empirically proven that in the diagnosis of breast cancer, the suggested model has a superior mean performance compared to the GA-KDE model.

A gene selection method that comprised two stages was suggested by Elyasigomari et al., [17]. This was referred to as MRMR- Cuckoo Optimization Algorithm (COA) and Harmony Search (HS) (MRMR-COA-HS). A subset of relevant genes is selected in the first stage using the MRMR feature selection. These genes which have been chosen are supplied into the wrapper set up which blends a new heuristic COA-HS, using the SVM as a classifier. There are four microarray data sets across which this technique has been applied and the leave one out cross validation technique has been used for assessing the performance. When the performance of this technique was compared with the other evolutionary algorithms, it was found that this algorithm performed better as it selected fewer genes while still maintaining high accuracy of classification. These selected genes were further studied for their functionality; Finally, the outcomes confirmed the biological relevancy of the chosen genes to every type of cancer.

3 METHODOLOGY

For classifying objects that are delineated by several hundred attributes, machine learning

techniques are used often. Nonetheless, the amount of data that is needed for providing reliable analysis grows exponentially with the increase in the data dimensionality. Searching for a data projection into a smaller number of variables (or features) that can preserve as much information as possible is a popular approach to this high dimensional dataset problem. A critical step in data mining is feature selection which is used across domains such as genetics, medicine and bioinformatics. In this section, the CFS, GA and HBMO algorithm based feature selection methods are discussed.

3.1 Dataset

The patients have been assigned to either of the subgroups that have been classified by Estrogen Receptors (ER) status. In order to select the markers, each subgroup has been analyzed separately. The allocation of patients in the ER-positive subgroup is randomly done into the training set of 80 patients and the testing sets of 129 patients. The ER-negative subgroup has been classified into training sets of 35 patients and testing sets of 42 patients [18].

3.2 Correlation Based Feature Selection (CFS)

The heuristic that is used to evaluate the value of a feature subset lies at the heart of the CFS algorithm .The viability of the individual features to predict the class labels along with the extent of their inter-correlation is taken into account by the heuristic. This is the hypothesis which forms the basis for this heuristic: Features that are highly correlated with the class, still, are uncorrelated to each other- is the attribute of a good feature subset [19].

The same principle holds good in the test theory for designing a composite test (the total or the mean of the single tests) to predict an extrinsic variable of interest. This scenario involves individual tests as "features" which evaluate the attributes corresponding to the variable of interest which is the class. For instance, a composite number of tests that evaluate a wide range of traits including the ability to learn, and, comprehend written material along with manual skill etc. will give a more accurate prediction of a person's success in the mechanics training course rather than, measuring a constricted scope of traits using individual tests.

Equation (1) formalizes the heuristic:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

Where $Merit_s$ is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{cf}}$ the average feature class correlation, and $\overline{r_{ff}}$ the average feature-feature inter-correlation. Equation 1 is, in fact, Pearson's correlation, where all variables have been standardized.

The numerator gives an indication of the prediction of a set of features while the denominator indicates their redundancy. As irrelevant features are poor predictors of class, these are handled by the heuristics. As the redundant features are highly correlated with one or more of the other features, these will be differentiated. The CFS , cannot, however, detect strongly interacting attributes such as in a parity problem as these attributes are treated autonomously. Yet, it has the ability to identify viable attributes under moderate levels of interaction [20].

Feature selection determines which of the features from the initial features set (possibly large) are to be included in the final subset and which are the ones that shouldn't be. There will be 2^n possible subsets for n features. These can be tries one by one which would be practically possible if only a small number of initial features are present.

In order to search for a feature subset in a reasonable time, several heuristic strategies like hill

climbing and best first are applied often. First a matrix of feature class is calculated by CFS followed by the correlation of features from the training data. This is followed by searching the feature subset space using a best first search. As best first search given superior results in some of the cases when compared to hill climbing, it has been used in final experiments. Starting with an empty feature set, the best first spawns all likely single feature expansions. Then, the selection of the subset that has the greatest evaluation is made and this is further expanded in the same way by appending single features. In case there is no improvement from expanding the subset, the search goes back to the next most viable subset (not expanded) and takes off from here. When there is enough time, the entire feature subset space is explored by the best first search; hence, it is not unusual to restrict the number of subsets that are expanded and have resulted in no improvement. When the search terminates, the best subset found is generated. CFS uses a stopping criterion of five consecutive fully expanded non-improving subsets [21].

CFS has a tendency to choose a "core" subset of features containing lesser redundancy as the correlations are globally estimated across all the training instances. It is also strongly predictive of the class. However, there are certain cases wherein a subset of features exist that can be forecast locally within a small region of the instance area. In such scenarios, there are some machine learning algorithms that have exploited the locally predictive features; under such circumstances the CFS has degraded their performance to some extent. In this work, the CFS version employed in experiments incorporates a heuristic for including locally predictive features and also avoiding the reintroduction of the redundancy. The searching of the feature subset space is followed by the examination of the remaining features that are unselected, one after the other, for determining their usefulness on a local scale and not on a global scale. If the correlation of a feature with a class is higher than the greater correlation between the feature and one of its features that has been already selected, a feature will be included into the subset.

3.3 Genetic Algorithm (GA) Based Feature Selection

The evolution theory of Darwin has inspired a clan of computational prototypes referred to as Genetic Algorithms (Gas). Darwin has recommended survival of the fittest. He has also stated that it is through the processes of reproduction, crossover and mutation that the survival of an organism can be maintained. The basic working mechanism of a GA is as follows: the algorithm begins with solutions set which is referred to as population. This set is represented by chromosomes. The solutions from one population can be utilized for generating a new population through the process of reproduction. The positive assumption made here is that the newer population is superior to the older one. This is why these algorithms are also known as optimistic search algorithms. The strategy for reproduction is in a manner that the chromosomes that demonstrate better solution to the target problem have better likelihood to reproduce when compared to those that demonstrate inferior solutions [22]. The best match is found by searching through a whole combination of parameters. For instance, to get a stronger, lighter and a better product overall, they can search through different combinations of material and designs and find the best combination of both. The basic principle of GA is [23]:

1) **Initialization**: This is marked by the set that comprises teh entire population or all of the sample points. These sample points may in turn comprise database tablets or directly scenarios from the real life. The principle used later alters in accordance. In case of real time data, the entire process takes the shape of natural language parsing after which statistics is

deployed. In database tables, semantic analytics are type casted into statistical theorem using random number generators.

A binary vector of the complete set of features represents every individual. Whether a feature is a part of the current feature subset is represented by each bit in the binary vector. If the ith feature is present in the current feature subset, xi=1 , else xi=0. For instance, the complete feature set is composed of six features, including feature 1 to feature 6. A vector such as (1, 1, 0, 1, 1, 0) means a selected feature subset including feature 1, feature 2, feature 4, and feature 5.

2) **Fitness calculating: For ranking a certain** feature subset against all the other feature subsets, the optimality of the feature subset has been described by fitness. GA selects individuals (feature subsets) having better fitness to take part in crossover and mutation so that the consequent population gets generated. The first requirement of a better feature subset is that there should be a greater information gain associated with each feature with respect to the class; Secondly, the feature that has been chosen should well represent the conservation of the stem. Thirdly, redundant features should be cast away while selection is done [24].

3) **Selection:** Darwin's theory of the survival of the fittest forms the basis for the notion of selection. From the set obtained from the previous step, a subset is chosen. This is a technique of categorizing data. Data that is comprised may appear to be logically correlated to a specific time instance or otherwise. There may also be several sets wherein every one of them has specific domain data such as data that gives the shopping trend of the clients or the data that gives a notion about the customer grievances. For determining which of the feature subsets take part in crossover and mutation, selection operation is applied. It also determines which of the feature subsets can be carried forward to the next generation.

4) **Cross over/ recombination: The** demarcation of the generic theory of the GA is the principle of cross over. The gene which is a single member of a certain set is crossed with another one from a different set. This leads to the interchange of conduct and trends across various sets which leads to the logical relationship being created between the set, thereby decreasing the randomness among the sets. At any point in time, it is only between the two respective genes that the crossover is performed and all the others are untouched.

In crossover which is a genetic operator, two parents/individuals produce new individuals/offspring. Now, the roulette wheel method selects two feature subsets in the current population as Parent1 and Parent2. After randomly selecting a point P at first, the bits that exist before P in Parent1 and Parent2 are retained; the bits after P are exchanged for generating two novel offspring feature subsets. Secondly, two randomly selected points are referred to as P1 and P2. Maintaining the bits between P1 and P2 , the remaining bits are exchanged. These two strategies for crossover are used alternately in the process of iteration; this enhances the diversity of the feature subsets [25].

5) **Mutation:** Beforehand, the objective of mutation was te generation of gender diversity. Mutation is a very critical manner in which the individual traits can be maintained as in crossover, the properties and traits of the other genes

replace the individual traits (alleles); It may be possible that using a combination of different genes, important alleles may be generated by mutation. In the mutation technique, a random point P is chosen and its value is reversed. Meaning, if P=0, the value will be set to 1 and vice versa.

6) **Acceptance: Mutation** results in the generation of the offspring – not all offspring can be viable candidates for the nest iteration. Hence the process of elimination comes into play. This elimination involves the calculation of the percentage of permissible traits for one gene. This marks the threshold similar to the activation mark in the neural networks. The genes that cross the threshold form the new population. This process goes on until two successive level genes have a negligible amount of difference between them.

The genetic iteration process stops when subject to two situations: First is when there is no modification whatsoever in the entire population's fitness in the recent N iteration or the difference is lesser than a threshold meaning that the population evolves very slowly. Second situation is for completing the GA, the maximum number of iterations are set and when this condition is met, the process of iteration ceases.

The steps of GA progresses in the following manner [26]:

Step1. Create initial population of candidate solutions.

Step2. Using appropriate fitness function, every individual is assigned a fitness value.

Step3. Fitness is evaluated and parents are chosen.

Step4. Offspring are created using reproduction operators i.e. crossover, mutation and selection on parents.

Step5. By choosing the offspring that has the best fitness, new population is generated.

Step6. Steps 3, 4, 5 are repeated until a termination condition is met..

3.4 Proposed Honey Bee Mating Optimization (HBMO) Algorithm Based Feature Selection

Honey bees are gregarious species of insects that follow a hierarchical and structured social order ; they construct hives; There is one queen along with several drones and workers in each hive. The queen is bigger than any other bee in the hive as she is fed the "royal jelly" which is a milky white, jelly like substance. She can lay 500 eggs in a single day and her lifespan is about 5-6 years. Sperm is provided to the queen by the drones. The mating drones have bigger eyes compared to the other bees. The drones that remained in the hive will be driven out to die, at the end of the season and thus, their lifespan is restricted to about six months only [27].

Workers comprise females who cannot reproduce. Instead they perform maintenance and operation tasks such as construction of the hive, rearing the brood that may come from fertilized egg that may be future queens/workers/unfertilized eggs that represent future drone bees. These workers also attend to the queen and the drones, clean the hives, regulate the temperatures , protect the hive when they are still young and gather nectar, pollen, water and some sticky plant resins that are used in the construction of the hives when they grow older. The workers born in the fall continue living until the following spring while those born early in the season live up to six weeks.

There are five main steps in the development of the HBMO algorithm, as described below.

Step 1: Mating flight of queen bees with drones

This process of marriage involves the mating of the queens with the drones whilst their mating flights in the air , far away from their hives. This process begins with the queen performing a waggle dance and being followed by the drones during the mating flight and culminating into airborne mating. Every mating results in the sperm attaining the spermatheca and gathering there so as to form a colonial genetic pool.A large swarm of drones known as the drone comets pursue the queen when the mating takes place. The death of the drone signals the end of the mating process and the culmination of insemination. Thus, the queen mates multiple times but the drone only once after which it dies. This is one of the most spectacular features of mating among insects.

For feature selection the HBMO algorithm is used. In the beginning, the population of the honeybees that will configure the initial hive has to be chosen. Each and every bee is arbitrarily placed in the d-dimensional space as a candidate solution (in the feature selection problem d corresponds to the number of activated features). Finding a suitable mapping between feature selection problem solutions and the bees in the HBMO algorithm is one of the main issues in formulating a successful algorithm for the feature selection problem. Each candidate feature is mapped into a binary particle wherein the corresponding feature chosen is denoted by 1 and the one that is not chosen is denoted by 0.

The algorithm has a real coded string representing every member of the colony of bees like the queen, the drones and the workers. The candidate solution to the problem is represented by a chromosome; every gene in the chromosome represents a parameter of the candidate solution. At the beginning of the mating flight, the speed of every queen bee is initialized randomly. Then, some drones are generated arbitrarily. After computing their objective function, the best of them is chosen to be the first queen. The next step is undertaking the mating flight during which there is a decrease in the speed.A drone mates with a queen probabilistically using an annealing function as (2):

$$prob(Q,D) = \exp\left(\frac{-\Delta(f)}{S(t)}\right) \quad (2)$$

Where prob (Q, D) is the probability of adding the sperm of drone D to the spermatheca of queen Q (that is, the probability of a successful mating), $\Delta(f)$ is the absolute difference between the fitness of drone, f (D) and the fitness of queen, f (Q) and S (t) the speed of the queen at time t .

After each transition in space, the speed and energy of the queen decays according to the following equations (3 to 5):

$$S(t+1) = \alpha S(t) \quad (3)$$

$$E(t+1) = E(t) - \gamma \quad (4)$$

$$\gamma = 0.5\frac{E(t_0)}{M} \quad (5)$$

Where $\alpha \in [0, 1]$, E (t) is the energy of the queen at time t, γ is the amount of speed reduction after each transition and M is the maximum number of mating flight.

The drone's sperm is deposited in the queen's spermatheca, in case of a successful mating (the drone passing the probabilistic decision rule). This is akin to simulated annealing wherein after generating a random number, a drone passes the probabilistic decision rule in case his probabilistic function exceeds the random number. A list of strings that belong to the drone's chromosomes having passed the probabilistic rule simulate the spermatheca in developing the algorithm. When the spermatheca of the queen is full- i.e. when she has mated to the maximum possible extent, and the energy is drained or when the speed reaches the

lower bound, the stopping criterion for this mating flight of the queen is reached.

Step 2: Creation of new broods by the queen

Whenever the queen lays fertilized eggs, she arbitrarily gets back the sperm mixture present in the spermatheca for fertilization. As per the algorithm, the queens start to breed and new broods develop once all the mating flights have been completed. The queen's genotype is mixed with that of the drones using the crossover operator. For the number of broods that are actually required, the selection of the queen is proportional to her fitness and she mates with some sperm from her spermatheca. Being very akin to GA, the process has one difference which is that in the GA the offspring is produced from two parents while in the HBMO, the brood may contain genes from multiple drones- meaning there is no certain male parent. Four crossover operators (intermediate, single point, two point, scattered) were considered in this work. Sensitivity analysis on the type of crossover was performed to find the most effective one [29].

Step 3: Improvement of the broods' fitness by workers

The workers take care of the brood and supply them with the royal jelly which is the queen's special food by means of which her size becomes larger compared to the other bees in the hive. When broods are fed with the royal jelly, their fitness improves and they have the ability to become the next queen.In the algorithm, this functionality of workers is modelled by representing them with a heuristic which acts to improve and/or take care of a set of broods. For improvising the genotype of the broods, a set of different heuristics is demonstrated by the workers.

Step 4: Adaptation of the workers' fitness

This step does not exist in nature. In theory however, as a reult of the heuristic application to the brood, the rate at which a brood's genotype improvises defines the fitness function for every worker. Thus, at every iteration, the fitness function for every worker gets updated so that the workers are given many more chances and this improvises the genotype of the brood. Thus, the next iteration employs the workers as per their fitness function [30].

Step 5: Replacement of the least fit queen(s) with the fittest brood(s)

Either fertilized or unfertilized eggs can arise from the brood. The potential queen bees or the workers are represented by the fertilized eggs and the prospective drones are represented by the unfertilized eggs. Even though a worker cannot be replaced by the brood, heuristically, the worst queen can be replaced by the best brood. The remainder of the broods wouldn't perish; however, the worst of the broods will be replaced by the best ones or the elite ones. This triggers the updation of a list of drones in every mating flight and the exploitation is powered by this replacement. Till the time all of the allocated mating flights have been completed or the convergence condition has been met, the fresh mating flight continues.

4 RESULTS AND DISCUSSION

In this section, the without feature selection, with CFS feature selection, with GA based feature selection and with proposed HBMO algorithm methods are used. Tables 1 to 3 and figures 1 to 3 shows the classification accuracy, f measure and positive predictive value. The figure 4 shows the percentage of feature selected.

Table 1 Classification Accuracy for Proposed HBMO Algorithm

Techniques	Classification accuracy
without feature selection	83.04
With CFS Feature Selection	85.65

With GA based Feature Selection	91.3
With Proposed HBMO Algorithm	93.04

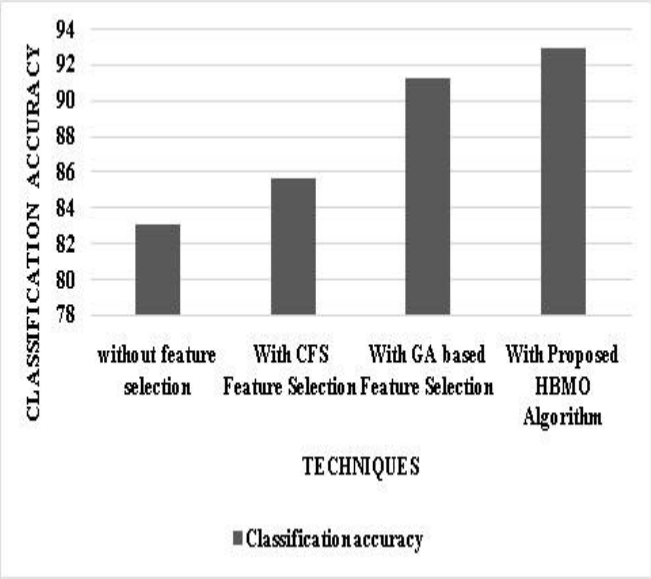

Figure 1 Classification Accuracy for Proposed HBMO Algorithm

From the figure 1, it can be observed that the with proposed HBMO algorithm has higher classification accuracy by 11.35% for without feature selection, by 8.27% for with CFS feature selection and by 1.88% for with GA based feature selection.

Table 2 F Measure for Proposed HBMO Algorithm

Techniques	F Measure
without feature selection	0.7879
With CFS Feature Selection	0.8239
With GA based Feature Selection	0.886
With Proposed HBMO Algorithm	0.9098

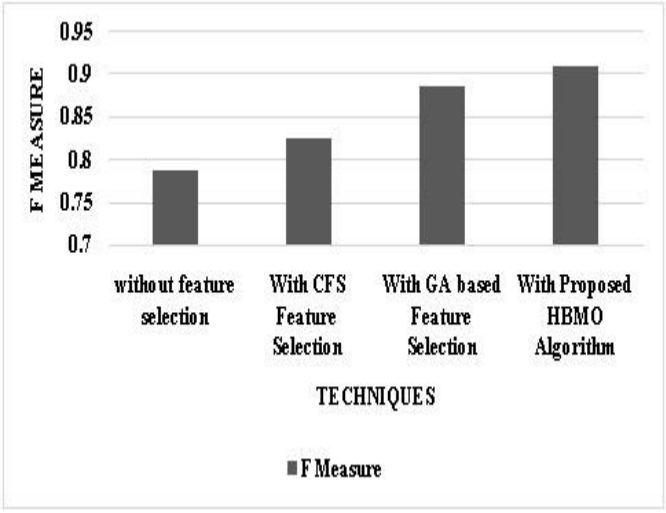

Figure 2 F Measure for Proposed HBMO Algorithm

From the figure 2, it can be observed that the with proposed HBMO algorithm has higher f measure by 14.36% for without feature selection, by 9.9% for with CFS feature selection and by 2.65% for with GA based feature selection.

Table 3 Positive Predictive Value for Proposed HBMO Algorithm

Techniques	Positive Predictive Value
without feature selection	0.77925
With CFS Feature Selection	0.81045
With GA based Feature Selection	0.8904
With Proposed HBMO Algorithm	0.9098

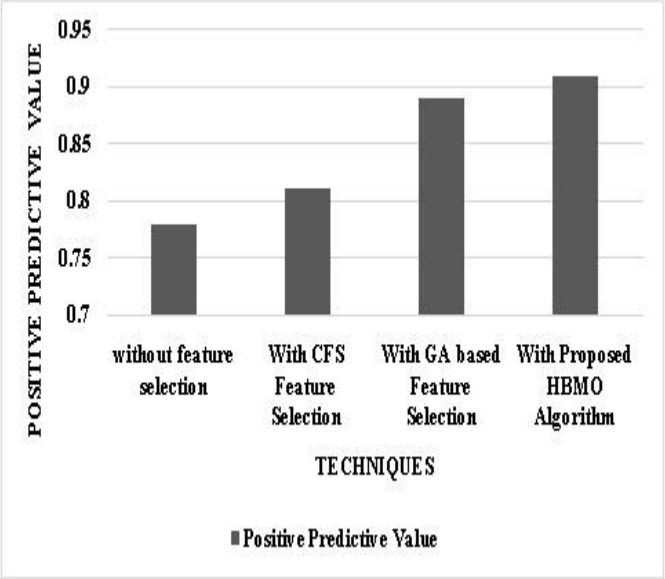

Figure 3 Positive Predictive Value for Proposed HBMO Algorithm

From the figure 3, it can be observed that the with proposed HBMO algorithm has higher positive predictive value by 15.45% for without feature selection, by 11.55% for with CFS feature selection and by 2.15% for with GA based feature selection.

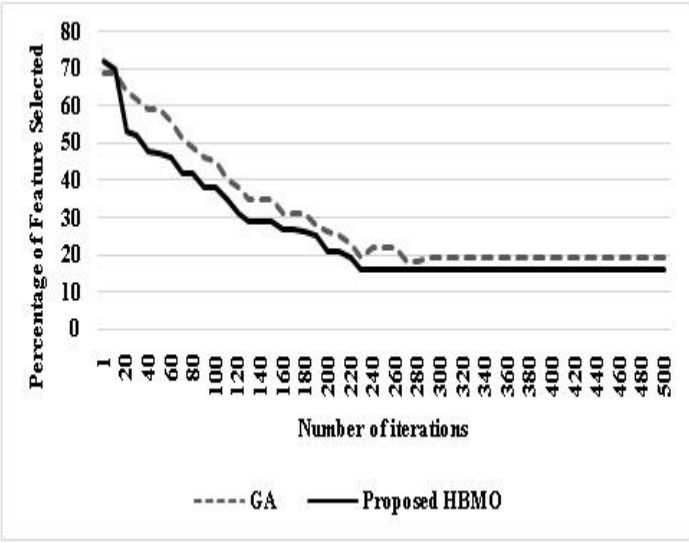

Figure 4 Percentage of Feature Selected

From the figure 4, it can be observed that the proposed HBMO has lower average percentage of feature selected by 16.14% compared for GA.

5 CONCLUSION

Data sets of large sizes can be referred to as big data. Processing the relevant and the non-redundant genes from the dataset is the most common challenge faced in bio informatics. Predicting and effective classification of genes will solve the issue of complex biological processes. Across fields like gene identification, cancer detection and disease diagnosis, prediction and treatment, there has been a widespread application of the microarray data. This can in turn trigger the development of medicines at a later stage. As the sample size id extremely small and the data is of high dimensionality, the classification problem is time consuming. The running time is reduced and the precision of forecast improves when the feature selection is performed before classification. This technique uses the adaptive GA approach which copies the genetic processes observed in the nature which can be applied to the optimization problems. GA was originally used to select binary strings and a number of authors have discussed the use of GA in feature selection. The most popular algorithm that has its basis in the marriage process of the bees which leads to the mating of the queen in the hive is the HBMO. Assuming that the queen is a superior bee, a randomly selected heuristic is utilized for improvising the queen's genotype. Therefore, the genes that will be transmitted to the broods are fixed for each drone. Results show that the with proposed HBMO algorithm has higher classification accuracy by 11.35% for without feature selection, by 8.27% for with CFS feature selection and by 1.88% for with GA based feature selection.

REFERENCES

1. Thun, M. J., DeLancey, J. O., Center, M. M., Jemal, A., & Ward, E. M. (2009). The global burden of cancer: priorities for prevention. Carcinogenesis, 31(1), 100-110.
2. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray

datasets and applied feature selection methods. Information Sciences, 282, 111-135.
3. Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.
4. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
5. Zakir, J., Seymour, T., & Berg, K. (2015). big data analytics. Issues in Information Systems, 16(2).
6. Dagade, V., Lagali, M., Avadhani, S., &Kalekar, P. (2015). Big Data Weather Analytics Using Hadoop. International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN, 0976-1353.
7. Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, 86, 33-45.
8. Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods, 111, 21-31.
9. Zhao, L., Chen, Z., Hu, Y., Min, G., & Jiang, Z. (2016). Distributed feature selection for efficient economic big data analysis. IEEE Transactions on Big Data.
10. Kong, H., Lai, Z., Wang, X., & Liu, F. (2016). Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning. Neurocomputing, 177, 198-205.
11. Wan, C., & Freitas, A. A. (2017). An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features. Artificial Intelligence Review, 1-40.
12. Zou, Q., Zeng, J., Cao, L., & Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing, 173, 346-354.
13. Mohammadi, M., Noghabi, H. S., Hodtani, G. A., &Mashhadi, H. R. (2016). Robust and stable gene selection via maximum–minimum correntropy criterion. Genomics, 107(2), 83-87.
14. Dashtban, M., & Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, 109(2), 91-107.
15. Li, G., Cui, L., Fu, X., Wen, Z., Lu, N., & Lu, J. (2017). Artificial bee colony algorithm with gene recombination for numerical function optimization. Applied Soft Computing, 52, 146-159.
16. Sheikhpour, R., Sarram, M. A., & Sheikhpour, R. (2016). Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Applied Soft Computing, 40, 113-131.
17. Elyasigomari, V., Lee, D. A., Screen, H. R., & Shaheed, M. H. (2017). Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. Journal of Biomedical Informatics, 67, 11-20.
18. Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., ...&Jatkoe, T. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet, 365(9460), 671-679.
19. Karegowda, A. G., Manjunath, A. S., &Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and

correlation based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), 271-277.
20. Hall, M. A. (1999). Correlation-based feature selection for machine learning.
21. Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. Department of Computer Science, University of Waikato, Hamilton, New Zealand.
22. Jain, S. (2017). Mining Big Data using Genetic Algorithm. International Research Journal of Engineering and Technology (IRJET), 4 (7), 743-747.
23. Hasan, M. (2014). Genetic Algorithm and its application to Big data Analysis. International Journal of Scientific & Engineering Research, 5(1), 1991-1996.
24. Xuan, P., Guo, M. Z., Wang, J., Wang, C. Y., Liu, X. Y., & Liu, Y. (2011). Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. Genet Mol Res, 10, 588-603.
25. Hong, S. S., Lee, W., & Han, M. M. (2015). The feature selection method based on genetic algorithm for efficient of text clustering and text classification. International Journal of Advances in Soft Computing & Its Applications, 7(1).
26. Hans, N., Mahajan, S., &Omkar, S. (2015). Big data clustering using genetic algorithm on hadoopmapreduce. International Journal of Scientific Technology Research, 4 (4), 58-62.
27. Marinakis, Y., Marinaki, M., &Matsatsinis, N. (2007, December). A hybrid clustering algorithm based on honey bees mating optimization and greedy randomized adaptive search procedure. In International Conference on Learning and Intelligent Optimization (pp. 138-152). Springer, Berlin, Heidelberg.
28. Karimi, S., Mostoufi, N., &Sotudeh-Gharebagh, R. (2014). Evaluating performance of Honey bee mating optimization. Journal of Optimization Theory and Applications, 160(3), 1020-1026.
29. Marinakis, Y., &Marinaki, M. (2011, July). A honey bees mating optimization algorithm for the open vehicle routing problem. In Proceedings of the 13th annual conference on Genetic and evolutionary computation (pp. 101-108). ACM.
30. Sabar, N. R., Ayob, M., Kendall, G., & Qu, R. (2012). A honey-bee mating optimization algorithm for educational timetabling problems. European Journal of Operational Research, 216(3), 533-543.