Firefly optimized k means clustering for gene selection

Magendiran N.[1]* and Selvarajan S.[2]
1. Associate Professor, Department of CSE, Paavai Engineering College, INDIA
2. Principal, Muthayammal College of Engineering, INDIA
*magendirannatarajanpec@paavai.edu.in

Abstract
The fundamental purpose of microarray or gene appearance data scrutiny is to identify the co-expressed genes as well as bright patterns. It also plays a vital role in investigating bioinformatics. Microarray technology can concurrently produce large quantity of microarray gene expression data for various samples, which permits effective analysis and diagnosis of breast cancer. It is not essential to have all the genes in the data set for classification as well as for diagnosis. While diagnosing breast cancer only informative genes are retained through gene selection process while redundant, inappropriate as well as noisy genes are discarded. This work proposes the pillar algorithm with K means as well as Weighted K Means (WKM) clustering algorithm and Firefly Algorithm (FA) based K means clustering.

Keywords: Gene selection, Weighted K Means (WKM) clustering algorithm, Firefly Algorithm (FA), Cluster Centre Initialization Algorithm (CCIA).

Introduction

The current generation has several techniques as well as scientific findings such as decision support system, image as well as scanning systems to assist clinicians in making suitable decisions. Since these techniques are expensive, remote areas are still lacking in such facilities. Consequently, the poor as well as the needy are deprived of such services. The doctor's knowledge and experience plays a major role in clinical decisions[1]. Many a times, these decisions are misleading as they are due to misinterpretations and it is becoming a very sensitive issue.

The most significant area of research is identifying genes that cause diseases. By identifying those attributes, it becomes easier to diagnose disease and to treat them appropriately. Illnesses like cancer are indicated through changes in expression value in few genes. For instance, healthy cells can turn into cancer cells through genetic mutation. Such change impact the expression levels of genes. Gene expressions refer to the procedure of copying genes' DNA sequences into RNA[2]. The expression level of gene denotes the basic quantity of copies of the genes' RNA generated in cells as well as is associated with the quantity of the associated proteins.

Several data mining algorithms which are built for micro-array genes expression data handles the clustering issue. Cluster analyses of genes expression data is proven as a beneficial tool to identify coexpressed genes. Every entrant in the environment is a metric of the appearance levels of a specific gene for a particular condition. Examining such datasets reveal genes of unfamiliar purposes as well as help to discover the functional relations among genes. Coexpressed genes may be collected into clusters on the basis of their appearance patterns of a gene. It is possible to achieve clustering through genes as well as models.

In gene grounded clustering, genes are conserved as objects and examples as attributes. In gene based clustering, the examples may be separated into homogeneous sets so that genes are observed as attributes whereas examples as objects. Clustering is relatively an unverified learning technique in which the substances under a subset of characteristics are clustered. Through clustering, individual objects are assigned into multiple groups[3]. Co-expressed genes with analogous expression designs may be clustered together that have same meanings.

The latest progress in DNA microarray technology allows obtaining gene expression profiles of tissue samples at comparatively lesser cost. Most of the scientists worldwide utilize the benefit of gene profiling to differentiate complicated biological circumstances as well as diseases [4]. Microarray methods, which are utilized in examining genome-wide, gene expression as well as genome mutation, assist scientists as well as physicians in comprehending the patho-physiological mechanisms used in diagnoses, prognoses as well as in selecting treatment plans.

Genes selection is a significant part for genes expression based tumour classification systems. The most beneficial aspect of microarray is that it can efficiently monitor the expression of large number of genes as well as offer biological information that are of great importance. Identifying the distinguished genes is very important and needs utmost attention [5]. Top ranking genes are thoroughly studied while the results support several research works in biology as well as medicine leading to many beneficial discoveries in cancer study. Medical diagnostic examinations that examine the existence of a given protein in serum can be attained through small subset of discriminant genes. Further, presence of supplementary feature will add on to the discriminating power of the genes. However, there are quite a few reasons to minimize the quantity of attributes to a sufficient minimum.

Clustering refers to the procedure of determining sets of objects so that the objects within a group are analogous to each other while it is dissimilar from the objects in other sets. An excellent clustering approach can create superior quality

clusters with high intracluster similitude as well as low intercluster similitude[6]. The factors that influence the quality of clustering results are the similitude metric employed by the approach as well as its implementation. It also depends on its capability to find a few or all of the unknown patterns.

Different types of clustering protocols are being proposed in accordance to the requirement. Clustering protocols may be sorted as hierarchical or partitional protocols on the basis of structures of abstraction. Hierarchical clustering protocols build a hierarchy of divisions, denoted as dendrograms wherein all partitions are nested in the partitions at the following hierarchical level. Partitional clustering protocols develop one partition, with pre-defined or predicted amount of non-overlapped clusters of data for recovering natural sets in data.

Medical data mining is highly efficient to explore hidden patterns in datasets belonging to medical field. Clinical diagnosis uses such patterns. Among the available data mining tools, Neural Networks are generally used for making prediction for medical information[7]. BPN make use of the gradient based method wherein there is slow training process or chances of getting trapped in local minimum. It is indeed better to apply the standard optimization techniques like the as Genetic Algorithm (GAs), Particle swarm optimization (PSOA), Ant Colony optimization algorithm (ACOA) for finding network weights rather than utilizing gradient-based learning methods.

Therefore, this paper suggests the Firefly Algorithm (FA) optimized K means clustering for gene selection. Section 2 reviews the literature for the related works for gene selection. Section 3 illustrates the techniques of k means CCIA, pillar optimization etc which are employed for the proposed work. Section 4 details the experimental outcomes and Section 5 gives the conclusion to the proposed work.

Related Work

Ng & McLachlan[8] implemented a clustering-based approach that is capable of working on complete gene-expression profile as well as draw inferences on differential expressions by utilizing weighted contrast of mixed impacts. Using an actual gene-expression dataset, the suggested clustering-based method may offer a set of marker genes, which enhances the predictions of disease outputs. The simulation outputs are used to compare the proposed method with other prevalent methods.

Srivastava et al.,[9] calculated the performance of filters versus wrappers genes selection method using supervised classifiers on three popular public domain data sets namely Ovarian Cancer, Lymphomas and Leukemia. In the case optimum genes selection, ReliefF technique is employed as filter-based genes selection as well as arbitrary genes sub-set selection protocol is utilized as wrapper based genes selection. For classifications, various linear and ensemble classifiers are experimented for the functionalities. By analyzing the methods, it is possible to identify the ones suitable for time management as well as the ones that can provide high precision while handling selected dataset.

Du et al[10] suggested a technique that applies feature genes selection as well as classification with SVM for micro-array data of lung tissues. On basis of the suggested approach, feature genes can possibly find out in accordance to epsilon-support vector regression (epsilon-SVR) as well as selection ordered genes from all classes. Further, applied multiclass support vector classification (multiclass SVC), crossvalidation as well as variable searching techniques to obtain higher predictions classification accuracy with lesser processing time. In this approach, the efficient dimensions decrease to find features genes is given utmost importance. The outputs prove that the feature genes that the proposed method find out can possible obtain high prediction classification accuracy.

Gormez et al[11] examined the statistical bias as well as variance of simple but still a standard feature selection algorithms by employing popular cross-validation methods on genes discovering study to predict hyper-tension. The results prove that the chosen genes are distinct for various techniques as well as varying crossvalidation could be run for both one gene as well as gene sub-set selections.

Jingbo et al[12] suggested an efficient machine-learning technique derived from artificial neural networks (ANN), to measure the possibility of a protein in rice to be disease resistant or not. By means of feature reduction almost 30 significant features associated with disease-resistance were found. Through feature selection technique it is possible to reduce the number of features to about 92.86%. Following this, a feature reduced classifier is built. The precision of the new classifier attained is 100% in re-substitution test while 72.13% in Jackknife test.The Matthews's correlation coefficient has achieved 0.4419. Consequently, top 10 probable Xoo-resistant genes are discovered.

Kamal et al[13] suggested three filtering methods namely Higher Weight (HW), Differential Minority Repeat (DMR) as well as Balanced Minority Repeat (BMR) for identifying the genes that are appropriate to fatal illnesses for biased micro-array expressions data. Experiment comparison with conventional ReliefF technique on 5 micro-array data sets illustrate efficacy of suggested techniques in choosing useful genes from micro-array expressions data with biased example distribution.

Methodology

This work proposes the pillar algorithm with K means as well as Weighted K Means (WKM) clustering algorithm and Firefly Algorithm (FA) based K means clustering.

Pillar Algorithm: The pillar protocol is tough as well as better to select the primary centroids for clustering protocol

by placing each centroid separately in data distribution. The Pillar algorithm is stated as [14]:

Assume $X = \{x_i \mid i = 1,...,n\}$ is data, k is the number of clusters, $C = \{c_i \mid i = 1,....,k\}$ the initial centroid, $SX \subseteq X$ is the identification for X that was previously chosen, $DM = \{x_i \mid i = 1,...,n\}$ is the aggregated distance measure, $D = \{x_i \mid i = 1,...,n\}$ is the distance measure for each iteration and the grand mean of X.

This work uses pillar optimization algorithm for selecting primary centroids for efficient gene selection to shun from situating centroids smartly as various positions generate several outputs. All points are assigned to the cluster with the nearest centroid. Following this, the centroids belonging to the clusters are revised by taking the means of the data points of every cluster. Some of the data points tend to relocate from one cluster to another. Fresh centroids are once again processed while the data points are designated to relevant clusters. Later, the assignment as well as revising centroids is frequented until convergence criterion is satisfied. In the protocol, mostly Euclidean distance $d_{ij} = \|x_i - c_k\|^2$ is utilized for finding the distance among the data points as well as the centroid [15]. The objective function is given by $d_{ij} = \|x_i - c_k\|^2$. In which $\|x_i - c_k\|^2$ represents the selected distance metric between data points xi as well as the cluster centers; ck denotes the distance of n data points from the corresponding cluster centres.

Gene Selection based on K-Means Clustering: Cluster analysis is a beneficial tool to identify groups in data independent of the decision parameter. In cluster analysis, it is hard to estimate the optimal quantity of clusters. The primary detached in cluster investigation is to group substances which are comparable to a single cluster as well as discrete objects which are different from Transmission them to different groups [16]. The most important maximum current clustering attitudes is K-Means clustering procedure. They are sorted objects to a predefined amount of clusters, that is expected by user accept K clusters. The knowledge is to select chance cluster centers, unique for every group.

Algorithm

Involve: $D = \{d1, d2, d3...dn\}$ *Dataset*

K – *Amount of chosen clusters*

Confirm: *A usual of K clusters.*

Steps:

1. *Randomly select k data points after D as original centroids*;

2.*Replication*

Allocate both opinion di to the cluster which has the neighboring centroid;

Compute the novel mean for both cluster;

Till meeting standards is found out.

Several researchers apply K-Means clustering to examine the gene expression data. This approach treats all the samples with equal importance during clustering. However, a few of the genes samples might get over expressed resulting in over weightage in each sample in the clusters. Thus this work proposes the WKM algorithm to examine the gene expression profile of breast cancer dataset. The weighted k-means clustering is appropriate for higher dimension data. This is because it obtains differential weights for the object, which provides a relative measure of importance to each variable in their respective cluster. N denotes initial set of centroids and it is determined by employing pillar algorithm while all the features are grouped to the nearest centroids in accordance to a Euclidean distance measure. Then, Weights are evaluated for each sample within each cluster employing the $Sum \quad s_i = \sum_{j=1}^{n} c_{ij} \quad i = 1,2,...,k$ and $w_i = (s_i - c_{ij}) / s_i \quad where \quad j = 1,2,...,n$.

The weights are a measure of the relative importance of all the samples with regard to the quantity of the features to that cluster. These weights are incorporated into the distance function $d_{ij} = \|w_i * (x_i - c_k)\|^2$ and again these features are clustered to its nearby centroids in order to decrease the distance for highly significant features. Then, centroids are revised. The procedure is iterated until a particular amount of iterations are reached or existing centroid is matched with previous centroid. The objective function $d_{ij} = \|w_i * (x_i - c_k)\|^2$, in which $w_i = (s_i - c_{ij}) / s_i \quad j = 1,2,...,n$, $Sum\ s_i = \sum_{j=1}^{n} c_{ij} \ i = 1,2,...,k$, is a selected distance metric among a weighted data point xi as well as the cluster centre; ck denotes the distance of n data points from the corresponding cluster centres..

Cluster Centre Initialization Algorithm for Gene Classification: The distinct gene expression matrix acquired from a scanning procedure includes noise, missing values as well as systematic variation emerging from the simulation process. Data preprocessing is necessary prior to achieving any cluster analyses [18]. Certain issues related to statistics pre-processing are gaining importance. Such questions are not included in this prediction; an investigation of the problem of missing value estimation as well as the issue of data normalization is addressed.

Algorithm

Contribution: DS = {d1, d2... dn} // data set

KM // Amount of chosen clusters

Production: A usual of KMpreliminary centroids.

Steps:

1. Established DS= 1;

2. fixedAmount the overhead among each data set and all other data established in the set DS;

3. Discovery the contiguous pair of data set from the set Ds and procedure a data set Am ($1 \leq m \leq KM$) which comprises these two data set, Remove these two data set from the set DS;

4. Novelty the data set in DS that is neighboring to the data set Am, Add it to Am and delete it from DS;
5. Replication step 4 until the amount of data sets in Am stretches 0.75 * (n/K);
6. If mean< KM, formerly mean = mean+1, find additional pair of data sets from DS among which the detachment is the shortest, from other dataset set Am and remove them since DS, Go to step 4;
7. For both data set Am (1 ≤ mean ≤ KM) find the mathematics Mean of the trajectories of data sets in Am, these resources will be the preliminary centroids.

Firefly Algorithm (FA): In standard FA, every iteration deal with brighter firefly that attracts other less bright fireflies toward itself in maximization optimization issues. The fireflies generally fly ignoring the global optimum. Consequently, it reduces the ability of FA to find the global best. For removing the weakness of FA as well as to improve the collective motion of the fireflies, an altered FA is proposed.

In the proposed FA, global optimum is utilized for the movement of fireflies. In each of the iteration, the globally optimal firefly draws the other fireflies to itself [19]. In MFA, Cartesian distance is utilized to calculate the distance of fireflies to global optimum, similar to standard FA as:

$$r_{i,best} = \sqrt{\left(x_i - x_{gbest}\right)^2 + \left(y_i - y_{gbest}\right)^2}$$

But for fireflies' movements:

$$X_i = x_i + \beta_0 e^{-\gamma_{ij}^2}\left(x_j - x_i\right) + \alpha\left(rand - \frac{1}{2}\right)$$

$$x_i = x_i + \left(\beta_0 e^{-\gamma_{ij}^2}\left(x_j - x_i\right) + \beta_0 e^{-\gamma_{i,gbest}^2}\left(x_{gbest} - x_i\right)\right) + \alpha\left(rand - 1/2\right)$$

g_{best} Represent global optimal while x_{gbes} represent the coordinate of global optima respectively.

As stated earlier, KCM is a popular and an easy method to cluster data. In K-means, firstly k random cluster centres are defined; data vectors are designated to every cluster on the basis of Euclidean distance. Every data vector is contrasted with k centre of clusters. Following this it is allocated to the nearer centre while the cluster center is refined. For improving the precision of k-means protocol, it is initialized with optimal centres, processed by FA as:

$$\left\lfloor Z_{1,1}, Z_{1,2}, ..., Z_{1,d}, Z_{2,1}, Z_{2,2}, ..., Z_{2,d}, ..., Z_{K,1}, Z_{K,2}, ..., Z_{K,d} \right\rfloor$$

In the suggested technique, the clustering consists of two stages. Firstly, it initializes fireflies with arbitrary values. As data is D-dimensional with K clusters, all fireflies have K×D dimensions. The objective function that should be minimal, is Euclidean distance. The technique of FA must be repeated until the pre-defined iteration. Secondly, the k-means would set with the position of suitable firefly. The centres are refined by means of k-means clustering.

Results and Discussion

Experiments were conducted using Wisconsin (breast cancer) and GEO Breast Cancer dataset. The algorithms are evaluated for classification accuracy, Specificity, sensitivity, f measure, Dice coefficient, and fowles mallow index. Table 1 and 2 summarizes the results achieved for K Means clustering with CCIA algorithm, Pillar optimization and firefly algorithm for Wisconsin and GEO dataset respectively.

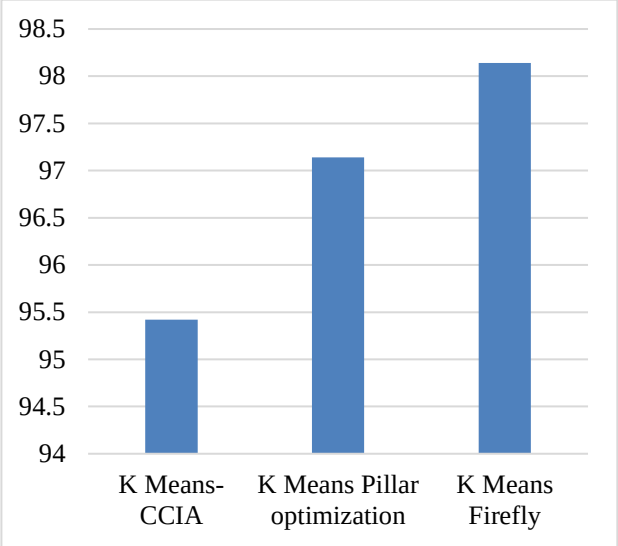

Figure 1: Classification Accuracy for optimized K Means Firefly (Wisconsin Dataset)

From the figure 1, it can be observed that the K Means Firefly has improved Classification Accuracy in Wisconsin dataset by 1.02% than K Means Pillar optimization and by 2.81% than K Means-CCIA.

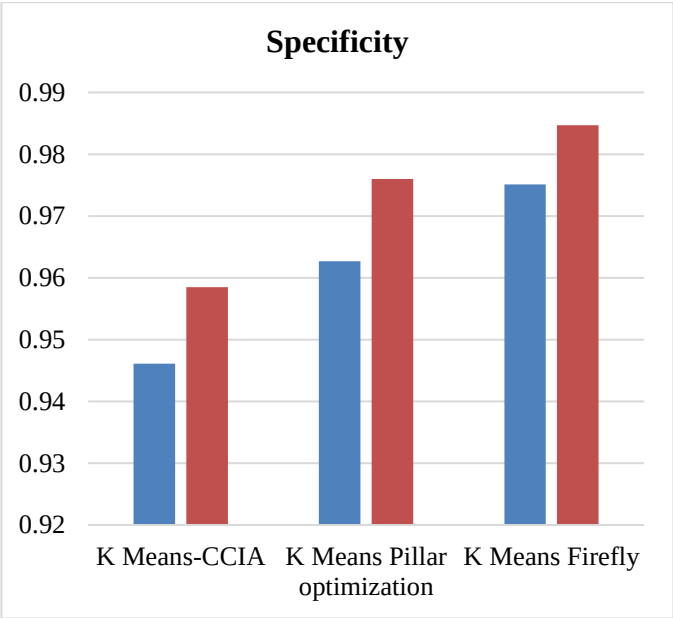

Figure 2: Specificity for optimized K Means Firefly (Wisconsin Dataset)

Table 1
Summary of Results for optimized K Means Firefly Clustering (Wisconsin Dataset)

	K Means-CCIA	K Means Pillar optimization	K Means Firefly
Classification accuracy	95.42	97.14	98.14
Specificity for Normal	0.9461	0.9627	0.9751
Specificity for Abnormal	0.9585	0.976	0.9847
Sensitivity for Normal	0.9461	0.9627	0.9751
Sensitivity for Abnormal	0.9585	0.976	0.9847
F measure for normal	0.9345	0.9587	0.9731
F measure for abnormal	0.9648	0.9781	0.9858
Fowles Mallow Index for Normal	0.9345	0.9587	0.9731
Fowles Mallow Index for Abnormal	0.9648	0.9781	0.9858
Dice Coefficient for Norm	0.9344	0.9587	0.9731
Dice Coefficient for Abnormal	0.9648	0.9781	0.9858

From the figure 2 it can be observed that the K Means Firefly for normal has improved Specificity in Wisconsin dataset by 1.28% than K Means Pillar optimization and by 3.02% than K Means-CCIA. The K Means Firefly for abnormal has improved Specificity in Wisconsin dataset by 0.89% than K Means Pillar optimization and by 2.70% than K Means-CCIA.

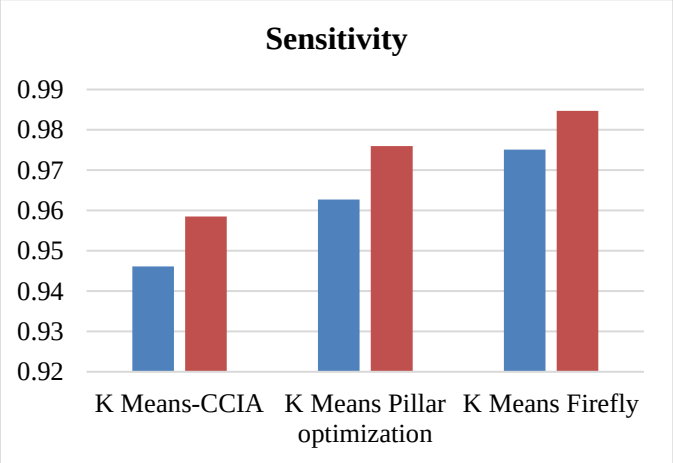

Figure 3: Sensitivity for optimized K Means Firefly Clustering (Wisconsin Dataset)

From the figure 3 it can be observed that the K Means Firefly for normal has improved Sensitivity in Wisconsin dataset by 1.28% than K Means Pillar optimization and by 3.02% than K Means-CCIA. The K Means Firefly for abnormal has improved Sensitivity in Wisconsin dataset by 0.89% than K Means Pillar optimization and by 2.70% than K Means-CCIA.

From the figure 4 it can be observed that the K Means Firefly for normal has improved F Measure in Wisconsin dataset by 1.49% than K Means Pillar optimization and by 4.05% than K Means-CCIA. The K Means Firefly for abnormal has improved F Measure in Wisconsin dataset by 0.78% than K Means Pillar optimization and by 2.15% than K Means-CCIA.

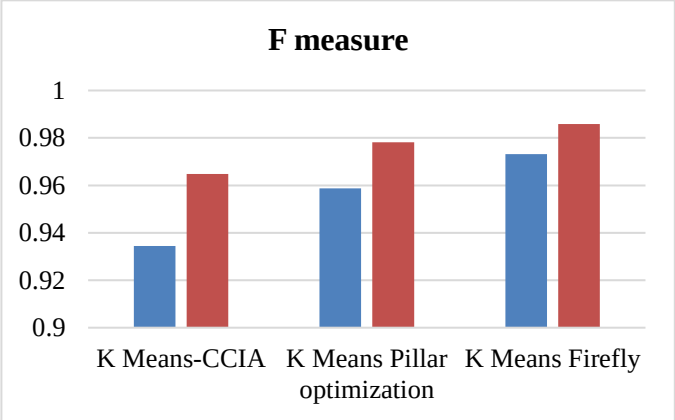

Figure 4: F Measure for optimized K Means Firefly Clustering (Wisconsin Dataset)

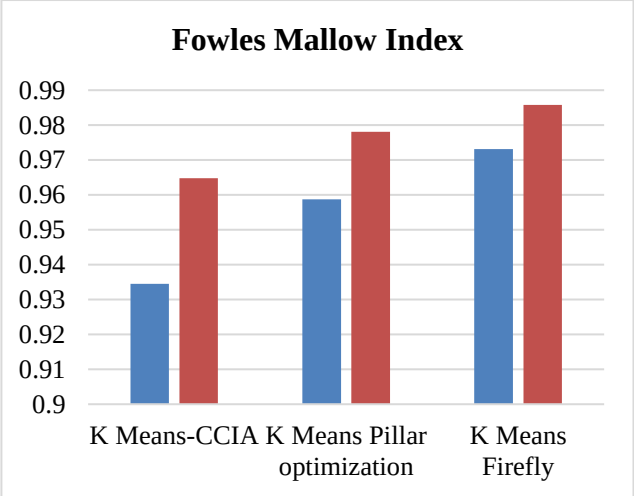

Figure 5: Fowles Mallow Index for optimized K Means Firefly Clustering (Wisconsin Dataset)

From the figure 5, it can be observed that the K Means Firefly for normal has improved Fowles Mallow Index in Wisconsin dataset by 1.49% than K Means Pillar optimization and by 4.05% than K Means-CCIA. The K Means Firefly for abnormal has improved Fowles Mallow Index in Wisconsin dataset by 0.78% than K Means Pillar optimization and by 2.15% than K Means-CCIA.

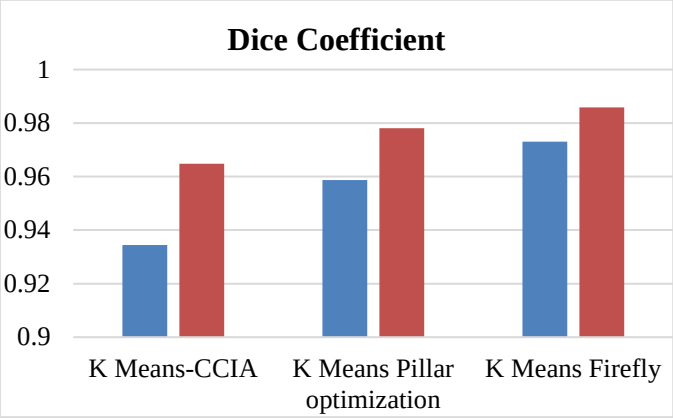

Figure 6: Dice Coefficient for optimized K Means Firefly Clustering (Wisconsin Dataset)

From the figure 6 it can be observed that the K Means Firefly for normal has improved Dice Coefficient in Wisconsin dataset by 1.49% than K Means Pillar optimization and by 4.06% than K Means-CCIA. The K Means Firefly for abnormal has improved Dice Coefficient in Wisconsin dataset by 0.78% than K Means Pillar optimization and by 2.15% than K Means-CCIA.

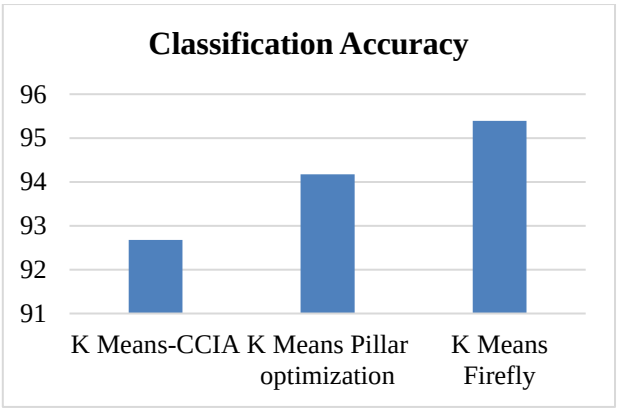

Figure 7: Classification Accuracy for optimized K Means Firefly (GEO Dataset)

From the figure 7 it can be observed that the K Means Firefly has improved Classification Accuracy in GEO dataset by 1.29% than K Means Pillar optimization and by 2.88% than K Means-CCIA.

From the figure 8 it can be observed that the K Means Firefly for ER Negative has improved Specificity in GEO dataset by 0.72% than K Means Pillar optimization and by 1.43% than K Means-CCIA. The K Means Firefly for ER Positive has improved Specificity in GEO dataset by 1.44% than K Means Pillar optimization and by 3.25% than K Means-CCIA.

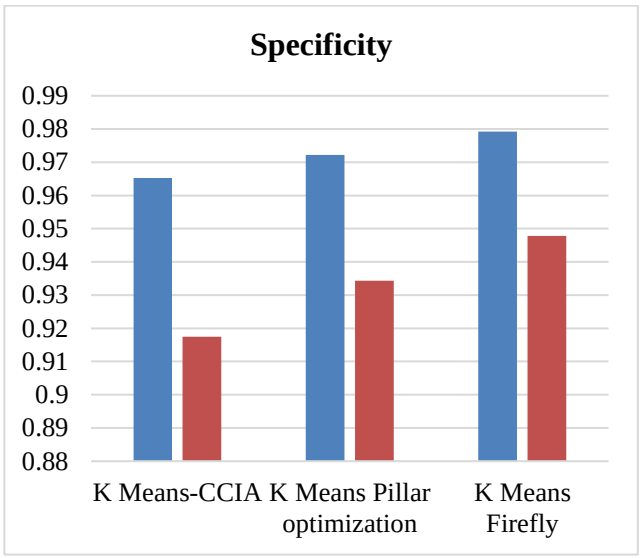

Figure 8: Specificity for optimized K Means Firefly Clustering (GEO Dataset)

From the figure 8 it can be observed that the K Means Firefly for ER Negative has improved Specificity in GEO dataset by 0.72% than K Means Pillar optimization and by 1.43% than K Means-CCIA. The K Means Firefly for ER Positive has improved Specificity in GEO dataset by 1.44% than K Means Pillar optimization and by 3.25% than K Means-CCIA.

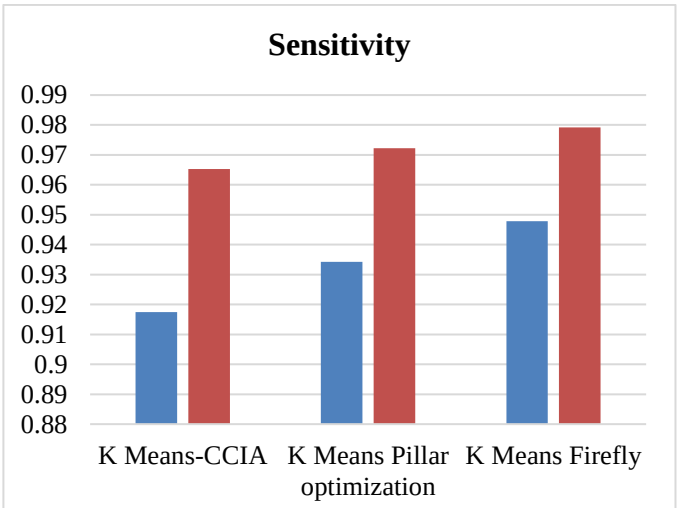

Figure 9: Sensitivity for optimized K Means Firefly Clustering (GEO Dataset)

From the figure 9 it can be observed that the K Means Firefly for ER Negative has improved Sensitivity in GEO dataset by 1.44% than K Means Pillar optimization and by 3.25% than K Means-CCIA. The K Means Firefly for ER Positive has improved Sensitivity in GEO dataset by 0.72% than K Means Pillar optimization and by 1.43% than K Means-CCIA.

Table 2
Summary of Results optimized K Means Firefly Clustering (GEO Dataset)

	K Means-CCIA	K Means Pillar optimization	K Means Firefly
Classification accuracy	92.68	94.17	95.39
Specificity for ER Negative	0.9653	0.9722	0.9792
Specificity for ER Positive	0.9175	0.9343	0.9478
Sensitivity for ER Negative	0.9175	0.9343	0.9478
Sensitivity for ER ER Positive	0.9653	0.9722	0.9792
F measure for ER Negative	0.9528	0.9627	0.9707
F measure for ER Positive	0.8374	0.8669	0.8924
Fowles Mallow Index for ER Negative	0.9535	0.9631	0.971
Fowles Mallow Index for ER Positive	0.8448	0.872	0.896
Dice Coefficient for ER Negative	0.9528	0.9627	0.9707
Dice Coefficient for ER Positive	0.8373	0.8669	0.8924

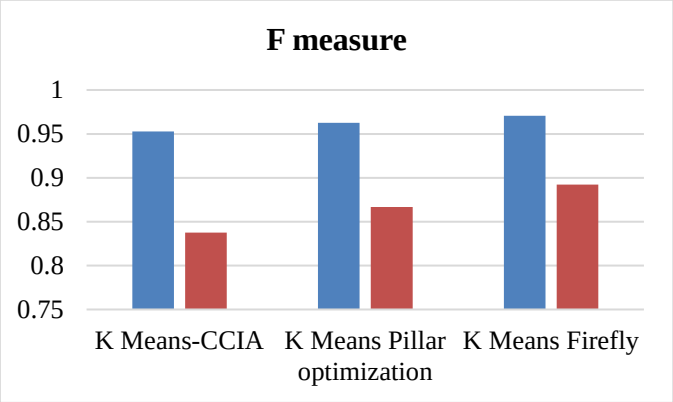

Figure 10: F Measure for optimized K Means Firefly Clustering (GEO Dataset)

From the figure 10 it can be observed that the K Means Firefly for ER Negative has improved F Measure in GEO dataset by 0.83% than K Means Pillar optimization and by 1.86% than K Means-CCIA. The K Means Firefly for ER Positive has improved F Measure in GEO dataset by 2.90% than K Means Pillar optimization and by 6.36% than K Means-CCIA.

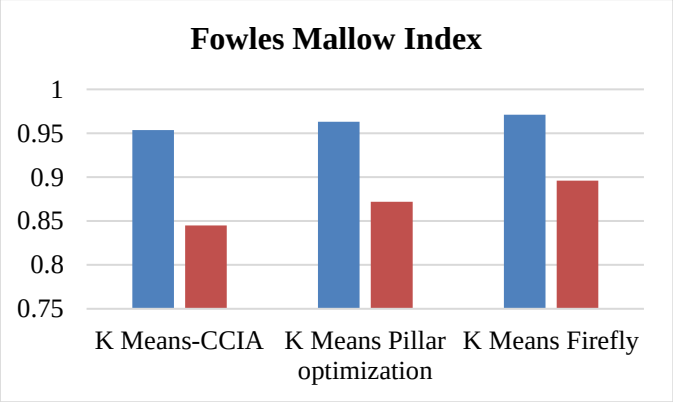

Figure 11: Fowles Mallow Index for optimized K Means Firefly Clustering (GEO Dataset)

From the figure 5.11 it can be observed that the K Means Firefly for ER Negative has improved Fowles Mallow Index in GEO dataset by 0.82% than K Means Pillar optimization and by 1.82% than K Means-CCIA. The K Means Firefly for ER Positive has improved Fowles Mallow Index in GEO dataset by 2.72% than K Means Pillar optimization and by 5.88% than K Means-CCIA.

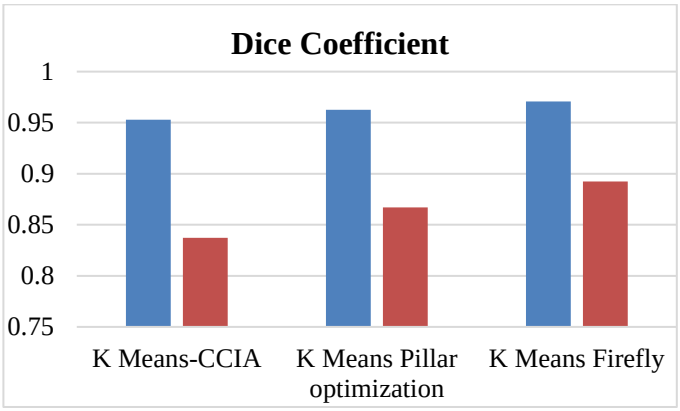

Figure 12: Dice Coefficient for optimized K Means Firefly Clustering (GEO Dataset)

From the figure 5.12 it can be observed that the K Means Firefly for ER Negative has improved Dice Coefficient in GEO dataset by 0.83% than K Means Pillar optimization and by 1.86% than K Means-CCIA. The K Means Firefly for ER Positive has improved Dice Coefficient in GEO dataset by 2.90% than K Means Pillar optimization and by 6.37% than K Means-CCIA.

Conclusion

This work proposes an innovative method derived from integration of clustering along with gap statistics, pillar optimization algorithm with K means as well as WKM algorithm to automatically choose more beneficial genes to deal the issues related to breast cancer diagnosis and

discovering drugs. It also proposes a modern hybrid method based on FA as well as KCM. Initially, FA is used to find optimum cluster centres while K-means is initialized with those centres to refine the centres. K Means firefly has performed better when compared to other methods in regard to classification accuracy, specificity, sensitivity as well as Fowles Mallow and Dice Index.

References

1. Ratnakar S., Rajeswari K. and Jacob R., Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes, *International Journal of Advanced Computational Engineering and Networking*, **1(2)**, 2320-2106 **(2013)**

2. Mohammadi A., Saraee M.H. and Salehi M., Identification of disease-causing genes using microarray data mining and Gene Ontology, *BMC medical genomics*, **4(1)**, 1-9 **(2011)**

3. Magendiran N. and Selvarajan S., Substantial Gene Selection in Disease Prediction based on Cluster Centre Initialization Algorithm, *Asian Journal of Research in Social Sciences and Humanities*, **6(cs1)**, 258-266 **(2016)**

4. Asyali M.H., Colak D., Demirkaya O. and Inan M.S., Gene expression profile classification: a review, *Current Bioinformatics*, **1(1)**, 55-73 **(2006)**

5. Praba, S. and Santra A.K., Tumor Clustering and Gene Selection Techniques-A Survey, *International Journal of Computer Applications*, **57(2)**, 1-8 **(2012)**

6. Sathishkumar E.N., Thangavel K. and Chandrasekhar T., A novel approach for single gene selection using clustering and dimensionality reduction, *International Journal of Scientific & Engineering Research*, **4(5)**, 1540-1545 **(2013)**

7. Karegowda A.G., Manjunath A.S. and Jayaram M.A., Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes, *International Journal on Soft Computing*, **2(2)**, 15-23 **(2011)**

8. Ng S.K. and McLachlan G.J., Using cluster analysis to improve gene selection in the formation of discriminant rules for the prediction of disease outcomes, In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference, 267-272 **(2013)**

9. Srivastava B., Srivastava R. and Jangid M., Filter vs. Wrapper approach for optimum gene selection of high dimensional gene expression dataset: An analysis with cancer datasets, In High Performance Computing and Applications (ICHPCA), 2014 International Conference, 1-6 **(2014)**

10. Du S.H., Su S.F., Jeng J.T. and Hsiao C.C., Feature genes selection and classification with SVM for microarray data of lung tissue, In Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium, 1054-1058 **(2014)**

11. Gormez Z., Kursun O., Sertbas A., Aydin N. and Seker H., Statistical bias and variance of gene selection and cross validation methods: A case study on hypertension prediction, In Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, 616-619 **(2012)**

12. Jingbo X., Xuehai H., Feng S., Xiaohui N. and Silan Z., Prediction of disease-resistant gene by using artificial neural network, In Research Challenges in Computer Science, 2009, ICRCCS'09, International Conference, IEEE, 81-84 **(2009)**

13. Kamal A.H., Zhu X. and Narayanan R., Gene selection for microarray expression data with imbalanced sample distributions, In Bioinformatics, Systems Biology and Intelligent Computing, 2009, IJCBS'09, International Joint Conference, 3-9 **(2009)**

14. Barakbah A.R. and Kiyoki Y., A pillar algorithm for k-means optimization by distance maximization for initial centroid designation, In Computational Intelligence and Data Mining, 2009, CIDM'09, IEEE Symposium, 61-68 **(2009)**

15. Li J., Su H., Chen H. and Futscher B.W., Optimal search-based gene subset selection for gene array cancer classification, *IEEE Transactions on information technology in biomedicine*, **11(4)**, 398-405 **(2007)**

16. Kumar S.S. and Inbarani H.H., Web 2.0 social bookmark selection for tag clustering, In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference, 510-516 **(2013)**

17. Maji P. and Das C., Relevant and significant supervised gene clusters for microarray cancer classification, *IEEE transactions on nano-bioscience*, **11(2)**, 161-168 **(2012)**

18. Selvakumar S. and HannahInbarani H., Covering Rough Set Based Intelligent Clustering Approach for Social E-Learning Systems, *International Journal of Applied Engineering Research*, **10(20)**, 19505-19510 **(2015)**

19. Hassanzadeh T. and Meybodi M.R., A new hybrid approach for data clustering using firefly algorithm and K-means, In Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium, 007-011 **(2012)**.