Towards Framing an Integrated Model for Optimization and Clustering (IMOC) High-Dimensional Non-Linear DNA Data Processing using ACO and DENCLUE

R.Nandhakumar, Antony Selvadoss Thanamani

Abstract: ***The Gene Expression data contains important information about the biological reactions that are carried out in creatures that are based on the surroundings. For understanding those processes in a better manner, the hidden expressions from the high-dimensional non-linear data are to be processed effectively. Moreover, the intricacy of biological data increases the challenges in High-dimensional non-linear gene data processing. For handling the complications, incorporation of clustering techniques is employed for identifying the patterns appropriately. Furthermore, this makes clarity in gene functions, regulations, inherent expressions, categories from noisy data input. In this paper, an Integrated Model for Optimization and Clustering (IMOC) is developed for efficient gene data processing for cancer detection application. Two efficient algorithms are integrated in this work, Ant Colony Optimization for effectively determine the features of gene data and Density based Clustering (DENCLUE) is for clustering the DNA data based on the determined features. For evaluating the proposed model, the benchmark datasets such as DNA Microarray Data of Leukemia and DNA Microarray Data of Colon Cancer are used. Further, the results show that the proposed model outperforms the existing models in accuracy and efficiency rates.***

Keywords: Gene Expression Data, DNA microarray data, Ant Colony Optimization, Density Based Clustering, Integrated Model.

I. INTRODUCTION

Data mining operations are widely used to derive attractive patterns from the available large database using efficient computing techniques. Particularly, in data mining, clustering is an important task to be performed for grouping similar objects or elements [1]. Nevertheless, High-dimensional data mining has been a great challenge among others, in which the clusters are framed with random based subspaces with the overall data dimensions [2]. In the high dimensional data clustering models, the distance measure is directly proportional to the increase of dimensional rates in the input data. DNA gene expressions are considered here as the high dimensional data.

The processing with microarray makes the research persons to evaluate the clarity of large number of genes of a certain tissue model for accurately find the gene with disorders. For such kind of evaluations, a computational model is required to process the huge volume of gene data [3]. Identification of coherent patterns in the high dimensional data is the important task in the research field of bioinformatics. The general structure of DNA microarray is given in Figure 1. Here, for effective analysis of gene expression data, clustering model is incorporated.

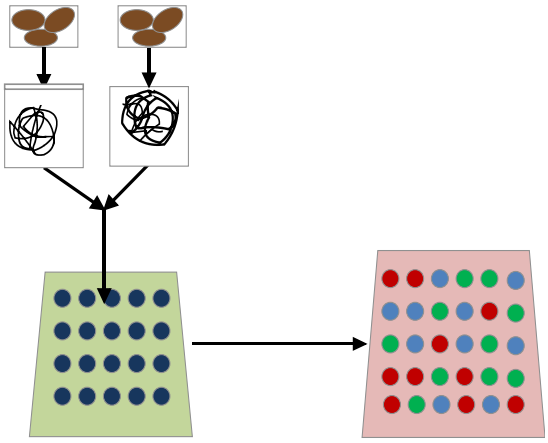

Figure 1: Model of DNA Microarray

For clustering in this model, density based clustering model is used. Moreover, an efficient Evolutionary algorithm called Ant Colony Optimization is also utilized for optimizing the surroundings. Moreover, the algorithm involved in handling, problems in data classification, data clustering [4] and routing issues [5]. This can be employed for effective determination of boundary of similar elements that are presented in the dataset.

Further, in data mining, Density based clustering is very familiar, in which, local cluster standards are used to define groups based on the data space region, having higher density of peak points. The random distribution of data points comprises the size and shape of the arbitrary region. A general process of determining the higher density regions is to find the higher density grid cells by dividing each dimensional space into non-connected grids. Hence, the algorithm can be categorized under grid based sub-space clustering. Initially subspace based clustering is done with the model called CLIQUE [6], for identifying the subspaces of highest dimensions. Another Clustering model is SUBCLU [7], is the enhanced model of DBSCAN.

In all clustering model of high dimensional non-linear data, the major function performed is compiling the data elements into groups called clusters based on their similarities. The clusters are

Revised Manuscript Received October 05, 2019

R.Nandhakumar, Assistant Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi-642001, India.

AntonySelvadoss Thanamani, Associate Professor & Head, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi-642001, India

defined based on the different considerations like fixed or distinct shapes and sizes. The sample clustering process is presented in Figure 2. The Density based Clustering (DENCLUE) model is a unique process of Kernel Density Estimation (KDE) [8], which is a non-parametric estimation pattern through which denser region points are determined. The DENCLUE model is more effective in clustering large volume of high dimensional, complex and noisy data.

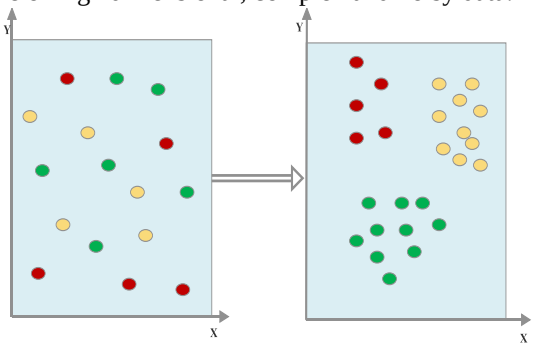

Figure 2: Clustering Pattern of DENCLUE

In this paper, the proposed Integrated Model for Optimization and Clustering (IMOC) combined the effectiveness of Ant Colony Optimization and Density based Clustering for determining the high dimensional non-linear gene data of DNA microarray data. The clustering issues can be effectively resolved using the functional behaviours of distinctive elements in the high dimensional data and clustering operation using DENCLUE is performed for identifying the cancer tissues that are presented in the sample data in accurate manner.

The remainder of this paper is framed as follows: Section 2 deliberates about the related work in high dimensional non-linear data processing. Section 3 explains about the work process of the proposed Integrated Model for Optimization and Clustering (IMOC). The results and comparative evaluations are provided in Section 4. Finally, the conclusion is given in Section 5 with some future enhancement directions.

II. RELATED WORKS

Several research works have been done to handle the issues on subspace clustering. In [9], CLIQUE (Clustering InQUEst) has been explained. It is a grid based clustering model, in which the repeated navigation is done with the bottom-up manner. The data space is divided into units with similar sizes based on parallel grids. The input parameters that are considered for partitions are the densities of data. As mentioned, the DNA microarray datasets are high-dimensional, non-linear; therefore, dimensionality reduction techniques are to be used effectively. The classification model for processing with DNA microarray data has been explained in [10] by reducing the features.

In the DNA microarray data, the number of elements in each category is not equal. In the scenario of high dimensional non-linear data, the region of interest may have fewer elements that are attributed to its label. The model that performed imbalanced DNA data processing was given in [11]. In that, a bias was generated for the majority classification. The imbalanced data could be made efficient with the process enforcement such as undersampling, oversampling techniques. The complexity level of processing increased when the rate of imbalanced data increases. Another work given in [12], it has been stated in the process about the reasons to be enhanced or corrected for increasing the performance of the processing of high dimensional non-linear data. The reasons stated by the authors are as follows,

- The machine learning techniques that are used for enhancing in such a manner to concentrate more on the majority classifications, when the model could not able to produce effective results in dealing with minority classification data.
- The region of interest determined with the fewer features, can made the model could not identify the real cause of the processing outcomes.

There are minimal numbers of researches in analyzing the similarity based evaluation in high dimensional non-linear data processing. For evaluating the similarities in distinct attributes of data, the model in paper [13] derived solution, which can be further applied for models with similarity measurements. A correlation based similarity measurement model has been developed in [14] for the feature analysis of data. Moreover, in [15], a stability derivation model has been given called Adjusted Stability Measure for employing on to the attribute subsets and to evaluate the subset stability in the process. The model specifically used to regulate the probability of similarity occurrences based on the auxiliary feature selection pattern.

In [16], Divisive Clustering Model was explained. It has been given in the paper that it was a top-down strategy model and the sample picture is given in Figure 3. It initiated with all the elements in a single cluster and divided till the final element has been reached. In this model, each cluster with N elements, there were (2N-1 -1) possible two subset partitions framed. Evaluating all the derived partitions was complicated and costly, specifically for gene expression data in divisive clustering.

In [17], Hierarchical based clustering has been explained in the processing of high dimensional gene expression data, which included the protein array and some other forms of bioinformatics. The model has been presented in Figure 4. The methodology has been effectively incorporated in cancer research and determining new categories of cancer [18] [19]. In [20], the hierarchical clustering methodology was used to define the variant sub categories of lung carcinoma. The results were to determine the profile of immunity based on two major categories here adeno-carcinoma and squamous.

The data from patients of ovarian tissue disorders has been evaluated in [21] for differentiating the malignant or early stages.

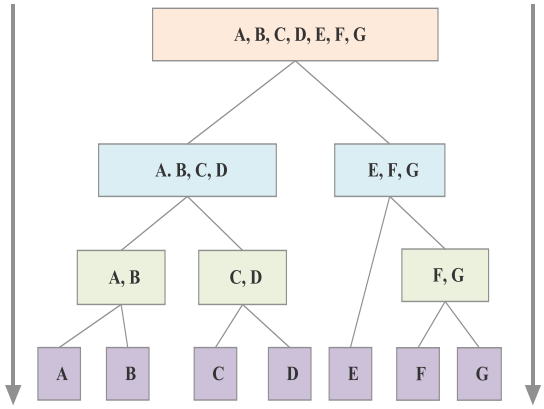

Figure 3: Representation of Divisive Clustering Model

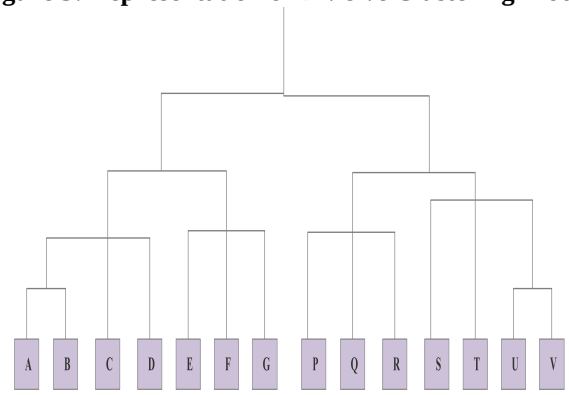

Figure 4: Hierarchical based Clustering Model

The Visualize Global Patterns in Gene Expression Data based model for processing high dimensional data has been described in [22]. Nevertheless, there are many issues to be cleared in hierarchical clustering. The major issues were adaptable to noise, high-dimensionality and so on. Additionally, when it was made to combine or divide clusters, it was not reframed or enhanced. The step wise cluster combining was evaluated in each levels based on the pair-wise separation instead considering global standards [23].

III. PROPOSED MODEL

This paper provides a novel method for determining cluster in DNA microarray gene data using the integrated concept of Ant Colony Optimization and Density based Clustering. The proposed Integrated Model for Optimization and Clustering (IMOC) is very efficient and appropriates in determination of clusters within the datasets. Moreover, two benchmark datasets are used in this process for evaluation.

A.Incorporation of ACO:

The process of ACO considers two things for deriving an efficient path selection for deriving efficient solutions for clustering problems. The chemical substances generated by the ants at the time of route determination travel are termed as pheromone, which can be used for providing communication between ants. The pheromone rate of productions reaches to a definite point, when the problem is solved or route is found successfully. Following, prior knowledge about the particular application is also used for reducing the costs and complications. Further, the model operates on two part of execution.

1. Transition Probability Criterion (TPC)
2. Factor Depiction

The first step involves in selecting the next point of movebased on TPC. The general formula for evaluating TPC is given as follows:

$$P_{st}(T)(TransitionProbability) = \begin{cases} \frac{pf_{st}^{\mu}(T)hc_{st}^{\mu}(T)}{\sum_{l\epsilon N_S} pf_{st}^{\mu}(T)hc_{st}^{\sigma}(T)}, & ift \in N_s \\ 0, & ift \notin N_s \end{cases} \quad (1)$$

where, '$P_{st}(T)$' is the TPC that the ant travels from source 's' to destination 't' at an instant T. 'N_s' is denoted as the optional next hop file of point 's', '$pf_{st}^{\mu}(T)$' denotes the pheromone focal-point on the travel space at time T with 'μ'as the pheromone analytical factor. And, hc_{st} is the heuristic content on the travel space at time T that specifically denotes the relationshipcondition of the travel space of ants or route determination cost. Moreover, the pheromone analytical factor represents the assignment of ants in the process of path framing. The strength of path is directly proportional to the value of the pheromone analytical factor. It can also be denoted in ACO as the path followed by the transmitted packet has a large contribution on the determination of present path, then, eventually, the rest of ants travel in the same path. When pheromone analytical factor becomes 0, the local best selection is employed, through which the ant selects the path that is considered to be the best in current state. The symbol 'σ' denotes ordinary analytical factor, determined by the relativeimplication of heuristic content. Further, the value of ordinary analytical factor is directionally proportional to the travel space quality affects the path determination. In the formula given for Transition Probability Criterion in (1), the importance is allotted for the hc_{st} and $pf_{st}^{\mu}(T)$. Further, the weight parameters are fixed on the basis of the variant situations, based on the simulation analysis. When an antcompletes route determination and moves to next point in travel space, the pheromone rate of the selected path is updated in with respect to the following derivation.

$$pf_{st}(T+1) = (1-\tau)pf_{st}(T) + \gamma\Delta pf_{st}(T) \quad (2)$$

Where, $\Delta pf_{st}(T)$ is the pheromone quantity produced by an ant at T on the travel space (s,t); 'γ' is the declining rate of pheromone, represents that the pheromone evanishes over time and the rate of γlies between, $\gamma \in [0,1]$, hence, $(1-\gamma)$represents the residual value of pheromone.

B. Work Process of Density based Clustering in Proposed Model:

In this proposed model, Density based Clustering is used for effective processing with high-dimensional non-linear data. Moreover, the DENCLUEE model is a unique derivation of Kernel Density Estimation, Which is a non-parametric evaluation process for determining the points at denser regions of the gene space. Here, the clustering process is incorporated to cluster the DNA microarray data for detecting the cancer. Initially, the DENCLUEE process works with two parts called pre-clustering and clustering, given in Figure 5.

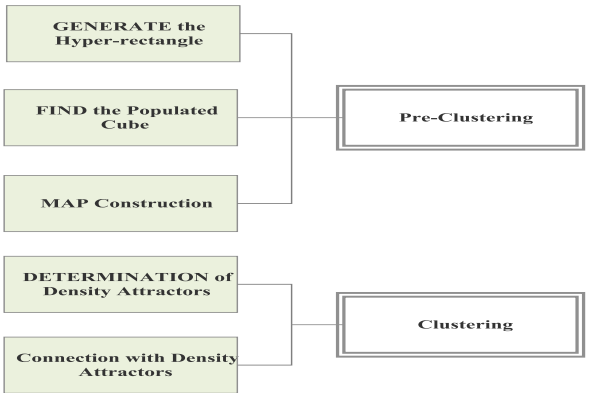

Figure 5: Operations in DENCLUEE

The process of pre-clustering includes construction of a hyper-rectangle map based on the dataset obtained. Moreover, the hyper-rectangle map is used for density function derivation. The hyper rectangle is framed based on the hyper cubes that are the most populated cubes from the datasets. Following, cluster determination is carried out from the highly denser or high-dimensional non-linear data that contains number of points surpasses a threshold λ given in factors and other adjacent number of populated cubes. Based on densities presented in the populated cubes on the sample data, the influence points are calculated between the cubes. From that, the density function is derived with the summation of the weighted points. There are many weight functions, which are formulated based on the distance between two points A and B. The weight function derived based on the points A and B are given in (3).

$$WeightFunction(WF) = exp^{\frac{-d(A,B)^2}{2\delta^2}} \quad (3)$$

From the above equation, d(A, B) is denoted the Euclidean Distance between A and B and δ denotes the radius of adjacent point that contains the A. Further, the density function is given in Equation 4.

$$DensityFunction(DF_{ds}) = \sum_{i=1}^{M} WF(A, A_i) \quad (4)$$

Where, 'ds' denotes the set of points in the obtained data sets and 'M' represents the cardinal points in the samples. Moreover, density factor is derived for each point for identifying the clusters using DENCLUEE. The density factor is specifically obtained from the local maximal value, which can be derived from the algorithm called hill climbing. The formation is presented in the following equation.

$$A = A^0, A^{i+1} = A^i = \sigma \frac{\nabla WF^{ds}(A^i)}{\|\nabla WF^{ds}(A^i)\|} \quad (5)$$

The final point of the computation is derived at $WF^{ds}(A^k) < WF^{ds}(A^{k+1})$ with $k \in M$. Here, $A^* = A^k$ is considered as the density factor. Using the points obtained from the derivation of density factors, a path is framed. The points are termed as the concerned points. Further, the clusters are formed based on the density factors and the concerned points.

C. Operations of Integrated Model for Optimization and Clustering (IMOC):

This paper integrates the efficiencies of two models, ACO and DENCLUEE for developing a novel model called IMOC. Based on the pheromone production and update principle of Ant Colony Optimization model, the density based clustering is carried out in the proposed model. Further, the steps involved in the proposed model are given as follows:

Step1. Feature Selection:

Since the input data is high dimensional non-linear microarray data, there are many distinctive features are presented, which may cause complications and errors in some cases. For solving that, dimensionality reduction is performed in the input DNA dataset by effectively selecting the features based on pheromone rate, produced by ants.

Step 2: Clustering Process:

This process is typically done with grouping up of similar data, based on that, the weight functions that are closer are to be clustered as in Density based clustering, given in the previous section.

Step 3: Boundary Determination:

Based on the clustering process accomplished at the previous step, the boundary between clusters are determined.

The pseudo code of the proposed model is given in Table 1.

```
//Feature Selection Process:
Begin
Fix threshold for feature_selection
Assign element_number
If (no. of feature >maximum No. of features)
{
For each element;
Create Test datasamples
Create Train datasamples
∀gene_data
Compute pheromone rate
}
If (data→paired-link) then
Calculate paired t-test
Else if
Calculate Unpaired t-test
Select the features with low pheromone rate
// Clustering Process
Generate hyper-rectangle
Calculate weight function
Calculate density function
//Boundary Determination
Add adjacent similar points
Compute new pheromone rate
Update pheromone rate
Evaluate Accuracy
Determine border to frame clusters
End
```

IV. RESULTS AND DISCUSSIONS

For evaluation purpose, the proposed IMOC model is evaluated and the results are compared with the existing models called CLIQUE and SUBCLU. The benchmark datasets that are used here are DNA Microarray Data of Leukemia and DNA Microarray Data of Colon Cancer. Further, the model is analyzed in MATLAB simulation platform. The factors that are considered for experimentation are number of ants, iteration rate

and evaporation rate of pheromone.

A. Dataset Description:

The data samples contain almost 7200 genes from 47 ALL called Actual Lymphoblastic Leukemia and 25 Actual Myeloid Leukemia, obtained from [24] DNA Microarray Data of Leukemia. The samples presented in the data are unpaired. The second dataset that is considered here, obtained from DNA Microarray Data of Colon Cancer [24], contains 2000 gene_samples from 62 patient records, among 40 are cancerous and 22 are non-cancerous. These samples are also independent of one another.

B.Comparative Evaluations:

For analyzing the proposed model, the accuracy rate of cancer diagnosis through the clustering model is more significant. The obtained results in accuracy estimation are presented in the Figure 6. By combining the two models called ACO and DENCLUE in the proposed IMOC, the boundary estimations are more accurate and thereby, the clustering of high dimensional non-linear gene expression data is appropriate. Hence, the detection of cancer from the input data sample is perfectly done from the input samples of DNA Microarray Data of Leukemia and DNA Microarray Data of Colon Cancer. It is clearly observed from the figure that the accuracy rate of the proposed model is 13% (in average) greater than the compared models.

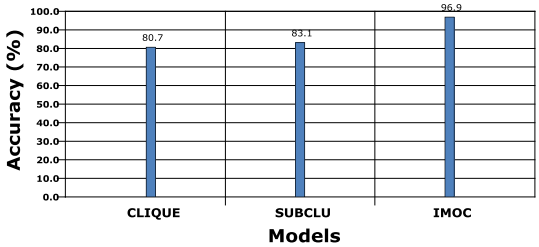

Figure 6: Accuracy Rate Evaluations of Compared Models

Moreover, in any process of clustering model, the performance can be evaluated based on the computations of sensitivity, specificity, precision and recall of the model, using the obtained true positive, true negative, false positive and false negative values. Based on those evaluations, the results are portrayed in Figure 7. It is obvious from the figure that the proposed model achieves better results than other models.

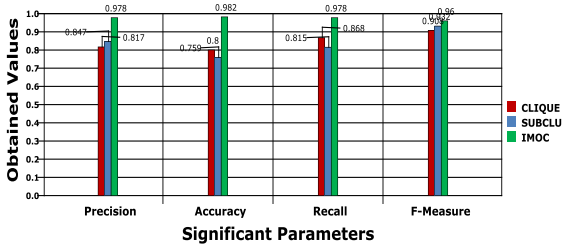

Figure 7: Factor based Evaluations

For an efficient processing model of high dimensional non-linear data, processing time is also to be considered. Because of the efficient incorporation of Evolutionary model and clustering model, the resultant clustering has been made with more accuracy in minimal time than other models. Here, the experimentation process is carried out with for the considered data sets. The results are provided in the Figure 8, for both data sets. Among models, for both datasets, the obtained processing time evaluated for the proposed model is lesser than the compared models called CLIQUE and SUBCLU.

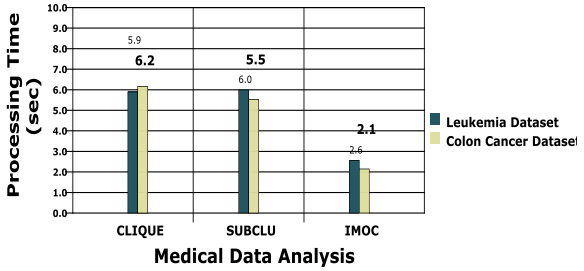

Figure 8: Processing Time Analysis between Models

V. CONCLUSIONS AND FUTURE WORK

This paper presents a novel model called Integrated Model for Optimization and Clustering for processing the high dimensional non-linear gene expression data effectively. The model combines the process of Ant Colony Optimization and Density based Clustering for effective group determination among the gene space of data samples. Moreover, the proposed model is evaluated using the available benchmark gene expression data and the results are compared with the existing approaches, for evidencing the effectiveness of the proposed model. Here, the overall processing time is effectively reduced using the ACO algorithm for cluster finding based on pheromone rate. Using, density based clustering; the similar elements are grouped under a single classification for improving appropriateness of boundary estimation, thereby increasing the accuracy of the prediction rate. Hence, the proposed IMOC model can achieve accurate results with minimal processing time any other processing and clustering models of high dimensional non-linear data.

In future, the work can be enhanced with some other mining models by incorporating fuzzy based logics in clustering formations.

REFERENCES

1. L. Yu and H. Liu, “Feature selection for high dimensional data: A fast correlation based filter solution”, In Proceedings of the Twentieth International Conference on Machine Learning, 2003, pp. 856-863.
2. R. Agrawal, J. Gehrke, D. Gunopulos and Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications”, In Proceedings of the SIGMOD, Vol. 27, Issue 2, 1998, pp. 94-105.
3. Sherlock G, “Analysis of large-scale Gene Expression Data”, Curr. Opin. Immunol.,Vol.12, 2000, pp. 201–205.
4. S. Gao, “A New Hybrid Ant Colony Algorithm for Clustering Problem”,International Sysposium on Intelligent Information Technology Application Workshops,Volume 1, 2008, pp.645-648.
5. T. Joan, P. Cristina, “An Ant-based Algorithm for Distributed RWA inOptical Burst Switching” , 12th International Conference on Transparent OpticalNetworks". S. Miguel, Azores: IEEE, 2009, pp. 1-4.
6. Siriphan W., “DNA Microarray Data Analysis Model usingClustering Algorithm for Disease Diagnosis”, Unpublished Master thesis,Prince of SongklaUniversit, 2008.
7. K. Kailing, H.P. Kriegel and P. Kroger, “Density-connected subspace clustering for high dimensional data,” In Proceedings of the 4th SIAM International Conference on Data Mining, Orlando, FL, 2004, pp. 46-257.
8. M. J. Zaki, W. MeiraJr, Data mining and analysis: fundamental concepts and algorithms, Cambridge University Press, 2014.
9. R Agrawal, J Gehrke, D Gunopulos, andP Raghavan. "Automatic Subspace Clustering

of HighDimensional Data for Data Mining Applications". InProc. ACM SIGMOD Int. Conf. on Management ofData, Seattle, WA, 1998.
10. I. Inza, P. Larraaga, R. Blanco, and A. J. Cerrolaza, "Filterversus wrapper gene selection approaches in dna microarraydomains," Artificial Intelligence in Medicine, vol. 31, no. 2, 2004, pp. 91 – 103.
11. R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," BMC Bioinformatics, 2010, pp. 523–539.
12. A. Al-Shahib, R. Breitling, and D. Gilbert, "Featureselection and the class imbalance problem in predictingprotein function from sequence," Applied Bioinformatics,vol. 4, no. 3, 2005, pp. 195–203.
13. L. I. Kuncheva, "A stability index for feature selection," in Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificialintelligence and applications. Anaheim, CA, USA: ACTA Press, 2007, pp. 390–395.
14. A. Kalousis, J. Prados, and M. Hilario, "Stability offeature selection algorithms: a study on high-dimensionalspaces," Knowledge and Information Systems, vol. 12,no. 1, 2006, pp. 95–116, Dec.
15. J. L. Lustgarten, V. Gopalakrishnan, and S. Visweswaran,"Measuring stability of feature selection in biomedicaldatasets," in AMIA 2009 Symposium Proceedings, 2009, pp.406–410.
16. Xu, R., and Wunsch, D., "Survey of clustering algorithms". IEEE Trans Neural Networks 16, 2005, pp.645–678.
17. Ramaswamy S., Ross K.N., Lander E.S., and Golub T.R. "A molecular signature of metastasisin primary solid tumors. Nat Genet. 33, 2003, pp.49–54.
18. Makretsov N.A., Huntsman D.G., Nielsen T.O., Yorida E., Peacock M., Cheang M.C.U., "Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significantgroups of breast carcinoma. Clin Cancer Res 10, 2004, pp.6143–6151.
19. Alizadeh A., Eisen M.B., Davis, R.E., MaC., Lossos I.S., Rosenwald A., "Distinct typesof diffuse large B-cell lymphoma identified by gene expression profiling". Nature 403, 2000, pp. 503–511.
20. Au N.H., Cheaug M., Huntsman D.G., Yorida A., Coldman A., and Elliott, W.M., "Evaluationof immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis:a tissue microarray study of 284 cases and 18 markers." J Pathol204, 2004, pp.101–109.
21. Mougeot J.L., Bahrani-mostafavi Z., Vachris J.C., Mckinney K.Q., GURLOV, S., Zhang, "Gene expression profiling of ovarian tissues for determination of molecular pathways reflective of tumorigenesis."J MolBiol358, 2006, pp.310–329.
22. Tseng G. "A comparative review of gene clustering in expression profile. Eighth International Conference onControl", Automation, Robotics and Vision (ICARCV). 2004, pp. 1320–1324.
23. Jiang D., Tang C., and Zhang A. "Cluster analysis for gene expression data: a survey". IEEE Trans Knowledgeand Data Eng 16, 2004, pp.1370–1386.
24. Kent Ridge Bio-medical Data Set. [On-line] Available:http://sdmc.lit.org.sh/GEDatasets/Datasets.html.

AUTHORS PROFILE

R. Nandhakumar , Assistant Professor in Computer Science, Nallamuthu Gounder Mahalingam College, undergoing his Ph.D in Data Mining. Published various papers under reputed journals. He is the coordinator in various activates in college like NAAC, ISO, etc..

Dr. Antony Selvadoss Thanamani, Associate Professor and Head, Research Department of Computer Science, He is Research supervisor for Ph.D. degree in Computer Science in the Bharathiar University, Dravidian University, etc.. He established Common Research centre, E-content studio, ISBN Nodal Agency at NGM College, Pollachi. He is Advisory Committee Member in National Conference on Advanced Computing. Editor in International Journal of Advanced Scientific Research, India . Editorial Board Member in International Journal of Advanced Research in Computer and Communication Engineering, India.