International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 2, April 2020, pp. 2060~2068
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i2.pp2060-2068

Novel modelling of clustering for enhanced classification performance on gene expression data

Sudha V.[1], Girijamma H. A.[2]
[1]Department of Information Science and Engineering, RNS Institute of Technology, Bengaluru, India
[2]Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru, India

Article Info

Article history:

Received May 27, 2019
Revised Oct 25, 2019
Accepted Nov 13, 2019

Keywords:

Accuracy
Classification
Clustering
Gene expression data
Genomics
Microarray data

ABSTRACT

Gene expression data is popularized for its capability to disclose various disease conditions. However, the conventional procedure to extract gene expression data itself incorporates various artifacts that offer challenges in diagnosis a complex disease indication and classification like cancer. Review of existing research approaches indicates that classification approaches are few to proven to be standard with respect to higher accuracy and applicable to gene expression data apart from unaddresed problems of computational complexity. Therefore, the proposed manuscript introduces a novel and simplified model capable using Graph Fourier Transform, Eigen Value and vector for offering better classification performance considering case study of microarray database, which is one typical example of gene expressiondata. The study outcome shows that proposed system offers comparatively better accuracy and reduced computational complexity with the existing clustering approaches.

Corresponding Author:

Sudha V.,
Department of Information Science and Engineering,
RNS Institute of Technology, Bengaluru, India.
Email: sudhavinayakam@gmail.com

1. INTRODUCTION

The area of genomic research has been consistently on demands due to various purposeful application e.g. forensics, medical examination, gene analysis, gene engineering, etc [1]. In this regards, microarray technology has made a considerable progress in last decade owing to its granularity of expressing the information within a gene [2]. There are presences of various types of artifacts in the actual gene expression data e.g. systematic fluctuation, missing value, noise, etc. Clustering approach also addresses this problem to some extent in order to facilitate better form of regulating gene and assists in investigating cellular function [3]. The clustering operation in gene can be carried out by grouping the gene as per their patterns of equivalent expression [4]. There are also good possibilities that genes with similar expression could fall under the same category of the cellular process. If there is a potential correlation associated with the patterns of gene expression that it is a direct indication of co-regulation of gene [5]. At present, it is believed that there are 10000-100000 genes present in one microarray data [6]. It could be even more for advance technologies of microarray data. The clustering process over gene expression data could be classified into sample-based and gene based [7]. The sample-based clustering scheme considers extracting features for genes and extracts objects from the sample, however, in gene-based clustering, it is vice-versa. Such samples can be used for representing a specific clinical condition. Irrespective of the differences, both the clustering approaches are used for searching the objects that has certain correlation with the some identified disease condition e.g. cancer. Normally, Euclidean distance is utilized for computing the proximity score between two objects present in gene expression data; however, they are not fit for addressing scaled patterns of objects in gene. Therefore, correlation-based methods are utilized for measuring the rate of

similarity. But still this method is also flawed as it cannot address the problems associated with outliers that results in false positives. The biggest problem with correlation-based approach is that if there is a common pattern between two different objects with single feature than discrete value of differences are not detected by this method. This method is definitely not robust for data or patterns with non-Gaussian distribution [8]. The clustering approach utilized for sample-based methods are supervised and unsupervised approach of selecting gene. As there are very less availability of the gene with potential information; therefore, selecting a gene with higher score of information is quite a challenging one. Apart from this there are other problems associated with sample-based clustering approach i.e. predefined number of cluster which is impractical and possess time complexity. Apart from the existing system, there are various other schemes to perform clustering approaches over gene expression data [9]. However, the problem is still unaddressed apart from various studies in existing system which is about selection of precise algorithm for particular genomic data. At present, there is no robust or full-proof approach to claim best clustering performance whereas it is found that only candidate algorithms are opted by researchers in order to perform comparative analysis. Therefore, this manuscript presents a discussion of a unique clustering algorithm which uses graph theory in order to construct a network of the entire significant object obtained from microarray data of gene expression structured logically. The paper gives a vivid explanation of the process undertaken in order to develop this mechanism of clustering and shows that proposed system offers better outcome. The organization of the paper is as follows: Section 1.1 discusses about the existing literatures where different techniques are discussed for detection schemes used in power transmission lines followed by discussion of research problems in Section 1.2 and proposed solution in 1.3. Section 2 discusses about algorithm implementation followed by discussion of result analysis in Section 3. Finally, the conclusive remarks are provided in Section 4.

- The background

There have been various approaches towards enhancing the clustering performance over the medical data [10]. The recent work carried out by Chen et al. [11] has introduced a network model for analyzing the redundant informationin the gene expression data. Identification of the specific form of cancer was carried out by Farouq et al. [12] over gene expression data. The approach uses a profiling mechanism for identification of disease condition using fusion-based mechanism. The work of Rosati et al. [13] has analyzed gene expression data where a hierarchical clustering approach mainly has been introduced using spatial information of the associated pattern. Existing system also finds that low rank clustering is another frequency used mechanism for analyzing complex medical data. The work of Liu et al. [14] has used regularization of hypergraph using a learning mechanism for performing subspace clustering. Usage of spectral clustering is another robust mechanism in order to explore overlapping region considering the case study of breast cancer (Luo et al. [15]). Sun et al. [16] have implemented a design of clustering mechanism on the basis of the affinity propagation where hybrid kernel system is introduced for obtaining effective precision. Deep learning is another frequently used clustering mechanism for investgating medical condition from gene expression data. Suo et al. [17] have used deep learning for clustering along with fuzzy c- means approach for carrying out clustering operation. Study towards involuntary training for the medical data is carried out by Xia et al. [18] where sub-space clustering approach using representation approach for low utilized ranks are harnessed. The work of Ahn et al. [19] have used time-series analysis integrated with clustering approach over gene expression data for solving the detection of specific forms of genes. Dominguez and Martin [20] have used a specific form of tools for computing similiarity score in order to generate a sophisticated network of genes. The model is claimed to offer reduced computational complexity in its clustering operation. Applying weights over the subspace is another strategy to improve clustering performance as seen in the work of Chen et al. [21]. Principal Component Analysis is another proven strategy to perform clustering over the gene expression data. The work of Feng et al. [22] has used Laplacian regularization process in order to optimize the clustering performance. Singular Value Decomposition using p-normalization approach as well as k-means clustering is another effective strategy to perform bimolecular clustering as seen in the work of Kong et al.[23]. Apart from this other schemes toward clustering operation are usage of ensemble classifier (Pratama [24]), Laplacian regularization with mix-norm (Wang et al. [25]), clustering on the basis of available information (Leale et al. [26]), matrix factorization (Li et al. [27]), random forest graph (Pouyan and Nourani [28]), integrated clustering using distance factor (Ushakov et al. [29]), and weighted consensus matrix (Wu et al. [30])]. Sudha V and Girijamma H A [31] has introduced a technique called SCDT for Gene study by using fuzzy cluster based closest neighbor categorization. Therefore, there are different variants of clustering scheme directed towards leveraging the clustering operation over gene expression data. All the mechanism has associated beneficial charecteristics as well as pitfalls too. The next section briefs of the open end research problems identified from this review.

- The research problem

The open end research issues associated with clustering approach in gene expression data are:

a. Existing clustering approach is based on implementing standard available clustering logic without any form of amendments on the top of it and thereby ignoring the associated issues.
b. Few approaches of clustering are actually found to be cost effective computation model towards assisting for solving complex disease classification problems.
c. There is no reported work on considering multi-dimensional data of the gene expression data for which reason existing clustering algorithm are less practical.
d. There was no much analysis of the computational complexity associated with the existing clustering approaches

Hence, the statement of the identified problem is "Obtaining a precise classification performance using cost-efficient design of clustering approach over gene expression data is quite challenging to achieve". The next section outlines the solution adopted to address this problem.

- The proposed solution

The proposed system is an extension of the prior clustering model where fuzzy logic has been used [31]. The model has focussed on framing up clustering framework; however, the proposed system extends this proposition by incorporating the classification operation using a cost effective modeling approach using graph theory. The schematic flow of the proposed system is highlighted in Figure 1.

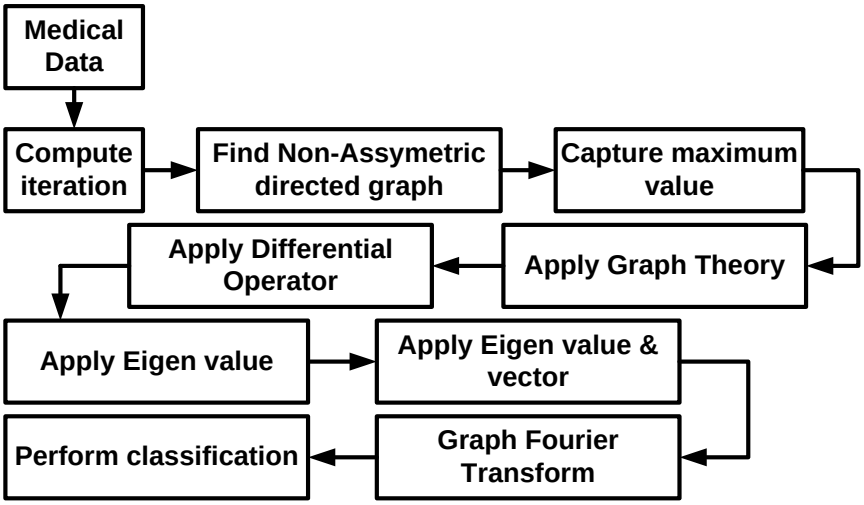

Figure 1. Schematic flow of proposed system

The proposed system considers a medical dataset in the form of complex gene expression data wih multiple discrete handlers. Each handler is subjected to defined set of iteration followed by applying non-assymetric directed graph. The system also considers maximum value of one of the handler in order to compute weight. In order to make the decision making system easier for classification, the proposed system constructs a network using graph theory followed by applying differential operator for better retrieval of classification vector. The proposed system also considers that there are good probabilityof fluctuations in the findings of the classification which can have an impact on the classification accuracy. Therefore, this research challenge is addressed by using Eigen value and Eigen vector which is capable of capturing the true information of any form of orientation. Finally, the proposed system applies Graph Fourier Transform along with conventional clustering approach of *k*-Nearest Neighborhood algorithm for efficient extraction of elite clusters associated with the condition of the cancer. Hypothetical clinical conventions of dual stages of cancers are then obtained as the outcome of the study representing the classification operation. In a nutshell, the proposed system offers a simplified approach which is capable of processing as well as analyzing the complex gene expression data for performing classification of the conditions of cancer. The next section illustrates about the system design and implementation of proposed classification process.

2. SYSTEM IMPLEMENTATION

The proposed algorithm basically harnesses the potential of graph partitioning system in order to perform classification of the diseases associated with medical data in the form of the logical matrix of variable dimensions. The research problem to solve in this state of implementation is associated with exploring a unique pattern of the complex medical data, which is in the form of microarray database. This section discusses about the strategies formulated for implementing the proposed algorithm following by illustrative discussion of the algorithm execution flow.

2.1. Adopted strategy for implementation

The *primary strategy* of the proposed system is to ensure that there is always certain form of ground truth values associated with the input of the gene expression data. Therefore, the dataset is considered in a unqiue way where there is an explcit flag information to clearly state the indication of type of cancer. It has to be understood that medical data in the form of image can be subjected for classification in the form of malilgnant and benign state based on some morphological condition of image (i.e. the input signal). However, this logic cannot be applied here as the signal is a form of gene expression data which is a logical matrix of elements 0 and 1. Hence the classification will be carried out with respect to type-1 and type-2 cancer which is at par with the numerical inputs of the dataset. This consideration is more practical and more realistic as it offers a true scale of numerical classification. The secondarystrategy of the proposed study is to ensure that the proposed system offers significant less computational overhead while perform classification operation and therefore, it is designed in such way that proposed system offers involvement of less iterative mechanism to evolve up in precision calculation. The proposed system uses K-Nearest Neighboring as clustering approach to obtained highly filtered outcome and has used graph foureir transform for better formation of the network to represent an unique clustering approach.

2.2. Execution flow of the classification algorithm

The proposed algorithm takes the input of the d (gene expression database) which is basically a form of complex medical data with an objective to perform cancer classification. The outcome of the algorithm is basically a v classification vector. The steps of the algorithm are as follows:

Algorithm for Classification
Input: d (gene expression database)
Output: v (Classification Vector)
Start
1. *init* $d=\{d_n\}$, where n is number of gene expression database
2. Obtain $h_n \rightarrow d_n$
3. gene_matrix$\rightarrow g_1(h_i)$
4. fluc_mat$\rightarrow g_2$(gene_matrix)
5. res_mat$\rightarrow g_3$(fluc_mat)
6. prec$\rightarrow g_4(h_i)$
7. v$\rightarrow$preci
End

The discussion of the execution flow of the proposed algorithm is as follows: The proposed algorithm considers multiple gene expression dataset d where n is the number of the type of the dataset. The study considers n=3, i.e. d_1, d_2, and d_3. The dataset d_1 represents signals of the genetic graph of the subject while d_2 represents the histology aspect of it. The dataset d_3 represents network of the gene which actually formulates gene vectors for making the computation easier. The significance of d_1 and d_2 are that d_1 offers a representation of the muted state of the subject's gene for flag value equivalent to 1 while the value of the flag is equivalent to 0 if it is non-muted. Similarly, d_2 dataset signifies Cancer State-I with a flag value of 1 while Cancer State-II is represented by flag value of 2 (Line-1). For simpler understanding it can be said that proposed algorithm obtains a simplified handlers h in the form of structure for the given input dataset d (Line-2). It will mean that h_1, h_2… h_n will be used for representing d_1, d_2, …, d_2 respectively. The next part of the algorithm is about accessing the gene structure using a function $g_1(x)$ with an input argument of the one handler h_i (Line-3). In the proposed system, the first handler is considered as signal of genetic graph of the subject. The working methodology of $g_1(x)$ is as follows: a) the dataset say d_1 is considered that consist of information associated with signal of genetic graph of the subject which is basically a logical matrix in nature with elements of 0 and 1 and specific dimension of m x n. This matrix is represented by its handler i.e. h_1. b) An iterative process *iter* is constructed if the sum of diagonal elements of all the elements of the handler h_1 is found to be non-zero. As the proposed system uses graph classification method therefore a directed graph

nsg is formed by accessing only the elements in the handler h_1 which are non-symmetric in nature. c) The algorithm than checks for the maximum value of the handler in order to obtain the weighted value *max_val*, d) The next process is to obtain the sum of all the non-zero elements of handler h_1, d) Finally, the primary and secondary fluctuation parameters are obtained. The primary fluctuation *prim_fluc* is obtained by subtracting diagnonal elements *diag_elem* with the handler h_1 while the secondary fluctuation *sec_fluc* is obtained by subtracting the handler matrix h_1 with all the diagonal elements within it h_1, e) The next task of the algorithm is to obtain the *network* information using graph classification on obtained secondary fluctuation *sec_fluc*, and f) differential operation *diff_op* is obtained by applying Laplacian operator to obtained *network*. Finally the step used in Line-3 results in a sparsity matrix for the given handler h_1.

The next part of the algorithm is to apply another function $g_2(x)$ which is responsible for evaluating the amount of fluctuation over the given matrix by obtaining the charecteritsic root (Line-4). The input argument of this function $g_2(x)$ are prior output arguments e.g. iteration *iter*, non-symmteric directed graph *nsg*, maximum value of the handler *max_val*, diagonal elements *diag_elem*, primary & secondary fluctuation *prim_fluc* and *sec_fluc*, network for classification *network*, and differential operation *diff_op*. Following operations are carried out in the following process of $g_2(x)$ function viz. a) The first process of this step of $g_2(x)$ function is to develop a matrix *up_prim_fluc* for updating the primary fluctuation value, b) the second process of this function is to obtain eigen value of updated primary function that results in complete matrix *full_mat* and diagonal matrix *diag_mat*, c) the third step is to obtain a sub-diagonal matrix *sub_dia_mat* from original diagonal matrix *dia_mat*, d) the fourth step is to obtain sequence *seq* value by sorting the absolute value of sub diagonal matrix *sub_dia_mat* in ascending order, e) this lead to generation of the updated version of the full and diagonal matrix with respect to order, f) the next step is to apply a conditional statement to check if the maximum value of the effective fluctuation is found less than cut-off value than the system obtains the final value of the updated primary fluctuation otherwise it flags inappropriate eigen decomposition outcome. g) Finally, all the shortlisted value of updated primary fluctuation are considered to obtain the final distinct value *distinct_val*. Therefore, this algorithm is primarily responsible for comparing the Eigen value of the signal with the cut-off variation to obtain the distinct value of fluctuation.

The next part of the algorithm implementation is all about applying another function $g_3(x)$ which takes the input argument of first handler h_1, second handler h_2, and third handler h_3. The flow of the underlying process are as follows: a) all the elements of the first handler h_1 is obtained and added up followed by obtaining the diagonal elements of it. The obtained diagonal elements are retained in a matrix *diag_elem*, b) The updated primary fluctuation *prim_fluc* is obtained in a similar way i.e subtracting the obtained diagonal matrix with the first handler h_1, c) obtain an eigen value as well as eigen vectors for primary fluctuation matrix with respect to length of the matrix. The obtained value is stored in complet matrix *full_mat* and inverse of this complete matrix is treated as diagonal matrix i.e. *diag_elem*, d) the next process is an iterative process for lower range of transformation where the third handler h_3 is multiplied with recently obtained matrix with diagonal element i.e. *diag_elem*. This operation will lead to generation of all possible transformation values explicitely for lower range of transformation to have a better control on the computational complexity. e) the final process is to shortlist transformation matrix of order 1 and 2 as *stage-1* and *stage-2* of cancer from the given gene expression dataset. The average value avg_1 and avg_2 are subsequently obtained for both *stage-1* and *stage-2*. vi) the total transform value *p* is obtained from obtaining absolute value of transformation matrix followed by summation of it. This operation also leads to generation of effective classification value *eff_class* by substracting both the obtained average and then divided by total transformed value.

The next part of the algorithm is about checking the level of accuracy using a new function $g_4(x)$ which takes the input arguments of third handler h_3, second handler h_2, transform matrix *trans*, and complete matrix *full_mat*. The operational steps of this algorithm are as follow: a) The first step is to apply *K*-nearest neighboring algorithm for clustering by considering the input argument of third handler h_3 and second handler h_2 which yields the outcome of primary optimal number x_1. b) this is followed by obtaining number of optimal outcome of clustering *x* and primary precision value *preci*, c) The next process is to generate primary optimal cluster O_1 for lower range of transformation, d) this process is followed by obtaining enhanced value of the transformation *enh_trans* by multiplying primary optimal cluster with graph fourier transform *trans*, e) the obtained transformation value enh_trans is than multiplied with complete matrix *full_mat* to obtain a scalar component *scal_com* of it. f) Similar process is continued to obtain secondary optimal number x_2 and secondary precision value *sec_preci*. Therefore, this stage of algorithm process is assists in computing the overall precision of the system as shown in Figure 2.

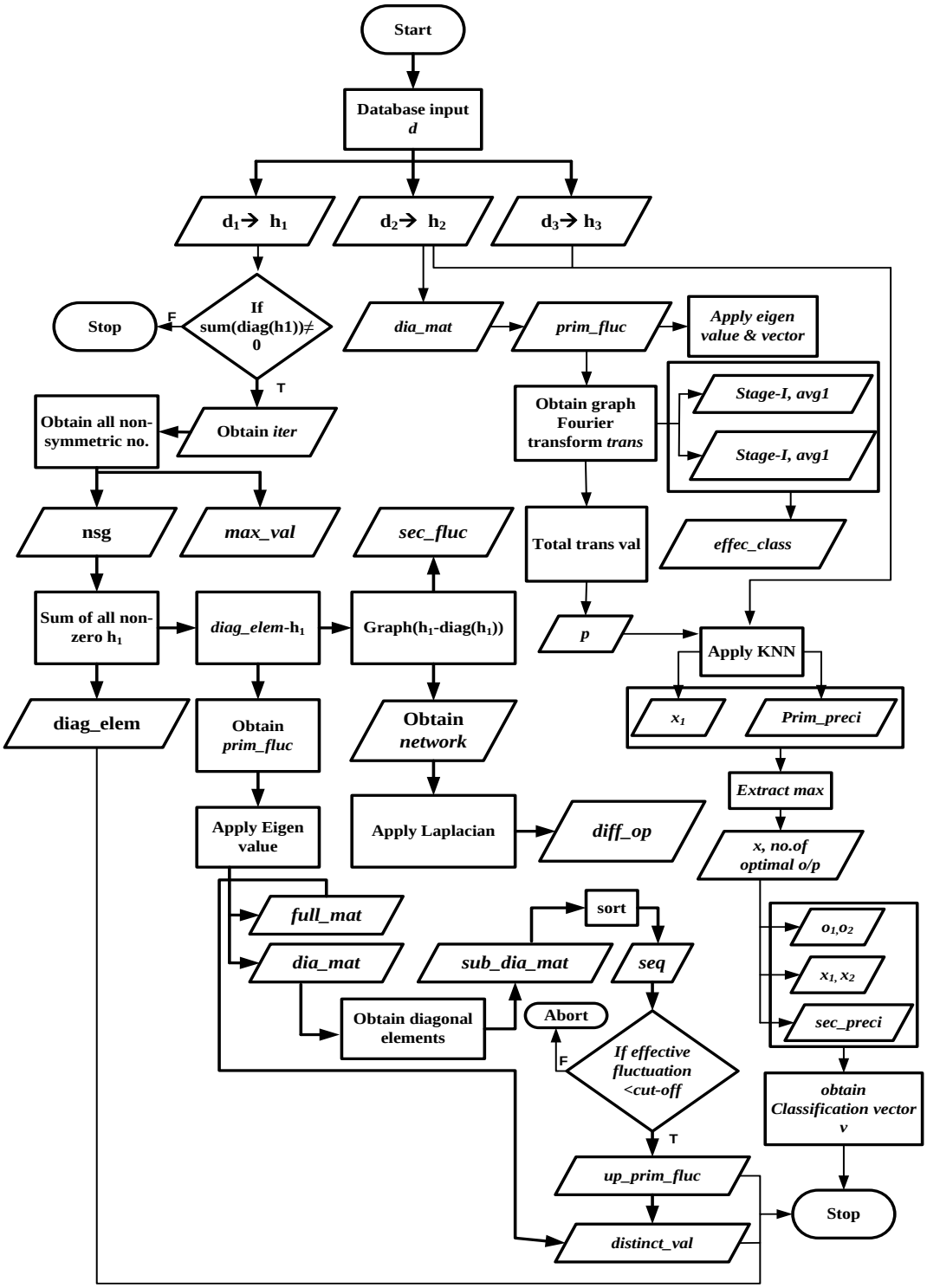

Figure 2. Proposed flow of process

3. RESULT ANALYSIS

The scripting of the implementation plan of proposed system is carried out in MATLAB considering normal system configuration of 4GM RAM and 2.20 GHz core-i3 processor in Windows system. The outcome of the study is compared with the existing approaches of frequently used clustering. The assessment is carried out over 100 iteration with respect to accuracy and processing time as performance parameters. The outcome of the proposed system shows that proposed system offers better accuracy as shown in Figure 3 and reduced processing time as shown in Figure 4 in comparison to existing clustering approach.

The prime reason behind this is that proposed system offer a comprehensive profiling scheme of the gene pattern which is in logical structure. Usage of differential operator assists in identifying the stages of the cancer with more accuracy. Apart from this, the usage of graph Fourier transform in a unique manner assists in forming better form of graph filter which is capable of identifying as well as mitigating the significant amount of artifacts or fluctuation obtained from the classification process. Apart from this the iteration considered for the analysis is basically a mapping of the k-value of the clusters which shows that with the increase of the cluster number, the proposed system do not have any negative impact on the performance parameters. Apart from this, none of the variables are found to retain more than 15% of the memory of complete graph processing. This fact will mean that proposed system also offer highly reduced spatial complexity apart from the reduced time complexity. Therefore, the proposed clustering proves not only to be cost effective but also a capability to perform classification of complex disease condition for a given set of gene expression data using multi-dimensional approach of clustering.

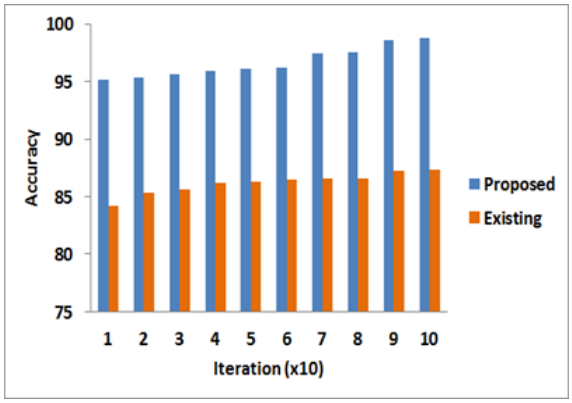

Figure 3. Comparative analysis of accuracy

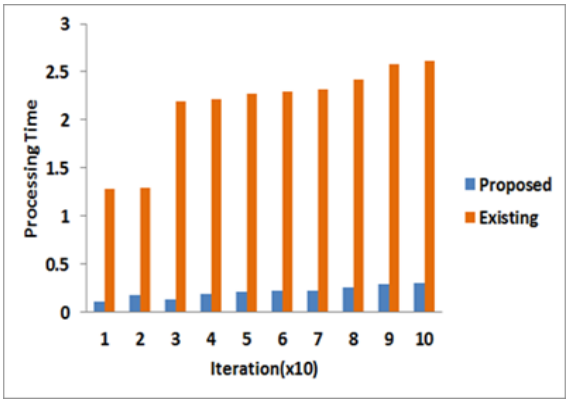

Figure 4. Comparative analysis of processing time

4. CONCLUSION

The significance of the clustering approach is realized in the proposed system where a multi-dimensional scheme of clustering has been introduced. The rationale of multi-dimensional clustering in proposed system is that it applies series of cluster formation using graph theory which facilitates in decision making. The proposed system also introduces a novel clustering mechanism which targets to achieve optimal number of clusters with highly reduced fluctutation degree in terms of its classification outcome. A significant contribution in the proposed approach could be noticed when it exhibits the balance of reduced computational complexity with increased accuracy of the classification for a given gene expression data.

REFERENCES

[1] Chris J. Myers, “Engineering Genetic Circuits”, CRC Press, 2016.
[2] Ho Nam Chang, Sang Yup Lee, Jens Nielsen, Gregory Stephanopoulos, “Emerging Areas in Bioengineering,” vol.1, John Wiley & Sons, 2018.
[3] Kim, Kwang Baek, and Doo Heon Song. "Intelligent Automatic Extraction of Canine Cataract Object with Dynamic Controlled Fuzzy C-Means based Quantization." *International Journal of Electrical & Computer Engineering*, vol. 8, no. 2, 2018.
[4] Lundmark A, Gerasimcik N, Båge T, Jemt A, "Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics", *PubMed-Science Report*, vol.8, iss.1, 2018.
[5] Omara, Hicham, Mohamed Lazaar, and Youness Tabii. "Effect of Feature Selection on Gene Expression Datasets Classification Accuracy." *International Journal of Electrical and Computer Engineering* 8, no. 5, 3194, 2018.
[6] Thomas Karn, "High-Throughput Gene Expression and Mutation Profiling: Current Methods and Future Perspectives," *PMC-Breast care*, vol.8, Iss.6, 2013.
[7] Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, "Clustering Algorithms: Their Application to Gene Expression Data", *PMC-Bioinform Biol Insights*, vol.10, 2016.
[8] Chung, Gwo Chin, Mohamad Yusoff Alias, and Jun Jiat Tiang. "Bit-error-rate Optimization for CDMA Ultra-wideband System Using Generalized Gaussian Approach." *International Journal of Electrical and Computer Engineering*, vol. 7, no. 5, pp: 2661-2673, 2017.

[9] Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, "Clustering Algorithms: Their Application to Gene Expression Data," *PMC-Bioinform Biol Insights*, vol.10,2016
[10] V. Sudha & H a, Girijamma, "Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.
[11] Chen, Xiang, Min Li, Ruiqing Zheng, Siyu Zhao, Fang-Xiang Wu, Yaohang Li, and Jianxin Wang. "A novel method of gene regulatory network structure inference from gene knock-out expression data," *Tsinghua Science and Technology*, vol. 24, no. 4, pp. 446-455, Aug. 2019.
[12] Farouq, Muhamed Wael, Wadii Boulila, Medhat Abdel-Aal, Amir Hussain, and Abdel Badeh Salem. "FGEP: A Multi-Stage Fusion-Based Approach for Gene Expression Profiling in Non-Small Lung Cancer of Non-Thermal Plasma Treatment" *IEEE Access*, 7:37141-37150, February 2019.
[13] Rosati, Paolo, Carmen A. Lupaşcu, and Domenico Tegolo. "Analysis of low-correlated spatial gene expression patterns: a clustering approach in the mouse brain data hosted in the Allen Brain Atlas" *IET Computer Vision*, vol. 12, no. 7, pp. 996-1006, 10 2018.
[14] J. Liu, Y. Cheng, X. Wang, X. Cui, Y. Kong and J. Du, "Low Rank Subspace Clustering via Discrete Constraint and Hypergraph Regularization for Tumor Molecular Pattern Discovery," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1500-1512, 1 Sept.-Oct. 2018.
[15] Luo, Jiawei, Ying Yin, Chu Pan, Gen Xiang, and Nguyen Hoang Tu. "Identifying functional modules in co-regulatory networks through overlapping spectral clustering." *IEEE transactions on nanobioscience* vol. 17, no. 2, pp: 134-144, 2018.
[16] L. Sun, R. Liu, J. Xu, S. Zhang and Y. Tian, "An Affinity Propagation Clustering Method Using Hybrid Kernel Function With LLE," in *IEEE Access*, vol. 6, pp. 68892-68909, 2018.
[17] Suo, Yina, Tingwei Liu, Xueyong Jia, and Fuxing Yu. "Application of clustering analysis in brain gene data based on deep learning." *IEEE Access*, vol. 7, pp: 2947-2956, 2019.
[18] Xia, Chun-Qiu, Ke Han, Yong Qi, Yang Zhang, and Dong-Jun Yu. "A self-training subspace clustering algorithm under low-rank representation for cancer classification on gene expression data." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 15, no. 4, pp: 1315-1324, 2018.
[19] Ahn, Hongryul, Heejoon Chae, Woosuk Jung, and Sun Kim. "Integration of heterogeneous time series gene expression data by clustering on time dimension." In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 332-335. IEEE, 2017.
[20] Gonzalez-Dominguez, Jorge, and Maria J. Martin. "MPIGeneNet: Parallel Calculation of Gene Co-Expression Networks on Multicore Clusters." *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 5 (2018): 1732-1737.
[21] Chen, Xiaojun, Joshua Zhexue Huang, Qingyao Wu, and Min Yang. "Subspace weighting co-clustering of gene expression data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 2, pp. 352-364, 1 March-April 2019.
[22] Feng, Chun-Mei, Ying-Lian Gao, Jin-Xing Liu, Chun-Hou Zheng, and Jiguo Yu. "PCA based on graph laplacian regularization and P-norm for gene selection and clustering." *IEEE transactions on nanobioscience*, vol. 16, no. 4, pp: 257-265, 2017.
[23] Kong, Xiang-Zhen, Jin-Xing Liu, Chun-Hou Zheng, Mi-Xiao Hou, and Juan Wang. "Robust and Efficient Biomolecular Clustering of Tumor Based on ${p} $-Norm Singular Value Decomposition." *IEEE transactions on nanobioscience*, vol. 16, no. 5, pp: 341-348, 2017.
[24] Pratama, M. Octaviano. "Kidney transplant classification with gene expression profiles using L1 feature selection ensemble classifier based on data clustering" In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 239-244. IEEE, 2017.
[25] Wang, Juan, Jin-Xing Liu, Chun-Hou Zheng, Ya-Xuan Wang, Xiang-Zhen Kong, and Chang-Gang Wen. "A Mixed-Norm Laplacian Regularized Low-Rank Representation Method for Tumor Samples Clustering." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* vol. 16, no. 1, pp: 172-182, 2019.
[26] Leale, Guillermo, Ariel Emilio Bayá, Diego H. Milone, Pablo M. Granitto, and Georgina Stegmayer. "Inferring unknown biological function by integration of GO annotations and gene expression data." *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 1, pp: 168-180, 2018.
[27] Li, Jianqiang, and Fei Wang. "Towards unsupervised gene selection: a matrix factorization framework." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol.14, no. 3, pp: 514-521, 2017.
[28] Pouyan, Maziyar Baran, and Mehrdad Nourani. "Clustering single-cell expression data using random forest graphs." *IEEE journal of biomedical and health informatics*, vol. 21, no. 4, pp: 1172-1181, 2017.
[29] Ushakov, Anton V., Xenia Klimentova, and Igor Vasilyev. "Bi-level and bi-objective p-median type problems for integrative clustering: application to analysis of cancer gene-expression and drug-response data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 1, pp: 46-59, 2018.
[30] Wu, Min, Le Ou-Yang, and Xiao-Li Li. "Protein complex detection via effective integration of base clustering solutions and co-complex affinity scores" *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 14, no. 3, pp: 733-739, 2017.
[31] V. Sudha and H. A. Girijamma, "Novel clustering of bigger and complex medical data by enhanced fuzzy logic structure," *2017 International Conference on Circuits, Controls, and Communications (CCUBE)*, Bangalore, pp. 131-135, 2017.

BIOGRAPHIES OF AUTHORS

SudhaV, currently working as Assistant Professor in the Department of Information Science and Engineering, RNS Institute of Technology, Bengaluru and having teaching experience of 11 year. Her research interests are in the field of Data mining, Data analytics and machine learning algorithms.

Dr. Girijamma H A, currently working as Professor in the Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru and having teaching experience of 23 year. Her research interests are in the field of automata, fuzzy logic and machine learning algorithms.