Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

MINING GENE EXPRESSION DATA BASED AFFINITY SEARCH CLUSTERING TECHNIQUE FOR CANCER THERAPEUTICS

Mr.U.Prakash[1], Ms.D.Shanmugapriya[2], Mr.Anguvigneshguru[3]

[1]*Information Technology, Sri Eshwar College of Engineering*
[2]*Information Technology, Sri Eshwar College of Engineering*
[3]*Information Technology, Accenture India*

Abstract— An appreciative towards genetics and epigenetic is necessary to deal with up by means of the pattern shift which is ongoing research in biomedical applications. Because of the development of new technologies, it has become very easier to collect the information of data regarding gene expression data in molecular ecology field. But analysis and examination of such type of data becomes very difficult since it consist of several genes and micro array dataset samples. So the clustering the gene expression becomes a valuable solution in the various fields of application such as business, medical knowledge, and economics. Conversely, traditional clustering algorithms methods for gene expression data produces less results ,since it is particularly designed for specific categories of gene samples .In order to conquer these problems in this work presents a affinity search based clustering methods for clustering gene expression dataset samples . Proposed work gene expression dataset samples are grouped based on measuring similarity value with predefined affinity threshold value. This clustering method the sub clusters are grouped based on the *k*-Medoids algorithm. This AFC consider each gene as objects and gene expression dataset samples as features. At the same time effectiveness is achieved higher than the traditional clustering method for gene expression dataset samples

Keywords- Gene expression data, gene therapy, epigenetic, next generation sequencing, clinic pathology, data mining, clustering and affinity search

I. INTRODUCTION

Approximately each and every one human genetic disease such as cancer and developmental abnormalities are distinguished through the occurrence of inherent variations. The detection and examination of gene expression prototype of a number of representation organisms characterize an interesting prospect to discover significant usual and irregular organic occurrence. DNA plays an essential responsibility at diverse bio-chemical procedure of livelihood organisms. Examining and analysis of DNA from larger dataset samples becomes a very difficult task in cancer succession and cancer investigate by means of high gene expression experimentation [1-3].

Clustering becomes one of the efficient methods to solve this in recent work some of the clustering methods have been also developed and introduced. Non-hierarchical method [4] is proposed for mining DNA sequences especially for oligonucleotide. In [5] initiate a pipeline technique designed for cis-Regulatory constituent forecast in mammalian genomes. A substitute schema is multicategory logit schema, to recognize original genes with the intention of demonstrate considerable correlations crosswise numerous period of prostate cancer detection have been introduced in [3]. In [6], another new schema is proposed to recognize human cancer genes the stage in a recessive way.

Cancer is a most important reason of each and every one the normal humanity and morbidities all the way through the world. Almost 13 percent of deaths are occurred during the

reason of the cancer [7]. Detection of the cancer and identification becomes very difficult task for prevention and cure them, although they are not moderately sufficient [8]. However, the detection and prevention of various types of cancer still becomes unsolved issues in today environment because of their various lifestyles.

Efficient data analysis approaches is required to detect the cause of cancer, so mining important data from dataset becomes very important for biological pathways and gene expression dataset samples. In this work presents an affinity search clustering method to mine and discover the information of gene biomarker. The biomarker genes are mined and identified based on measuring the similarity between gene expression dataset samples with predefined affinity threshold value to group cluster gene samples into k mediods cluster. The concentration of proposed clustering is to determine the results of oncogene dataset samples into normal and diseased sample. The organization of the paper is summarized as follows. Section 2 introduces existing methods designed for examination of gene expression data. In Section 3, study the working procedure affinity search clustering for gene expression data in cancer Therapeutics. In Section 4, the experimentation results of the methods have been measured and evaluated using various dataset for gene expression data. Section 5 provides a conclusion to the work.

II. RELATED WORK

Microarray-based proportional genomic hybridization method is introduced in [9] to perform mining for gene expression DNA dataset samples through high-resolution designed for the examination of human genome connected through actions deviation. A self adaptive and incremental neural network based classification schema is also introduced in this work to classify the lymphoma cancer samples into normal and disseminate large B-cell lymphoma (DLBCL) for cDNA microarrays dataset samples. Mean shift segmentation methods based clustering methods is introduced [10] for mining DNA microarray datasets samples.

In [11], introduction fuzzy relations schema is also introduced between breast cancer predictive reason and gene expression dataset samples. A recently considered data mining representation to accumulate microarray investigational information in a systematical association, and to give a well-organized technique designed for researchers to extract the database and inhabit it in a practical way intended for investigate development is characterized in [12]. Though, DNA microarray examination is extremely difficult appropriate to huge amount of genes and the noise with the intention of influence the complete procedure. Co-clustering techniques methods have been also proposed for mining the gene expression dataset samples based on the distance function and cluster quality procedures designed for integrating gene expression data. They co-clustering system usually group the gene expression dataset samples through comparable expression patterns [13]. To reduce patterns beginning the incorporated information, it is very important to work through classification rules, and clustering algorithms. It moreover focal point on choosing and eliminate irrelevant and difficult rules.

Gene-based clustering methods the number of genes is consider as objects and total number of gene dataset samples is considered as features, it is opposite to sample-based clustering [14]. The third category of clustering is subspace clustering, it performs clustering based on the class labels of the diseases, so it produces best clustering results when compare to other clustering methods. to perform clustering for above clustering methods similarity measure is important to decide which cluster data points is under the cluster or not ,various similarity heave been used in the recent work [15-16]. Among them all of the similarity metric correlation coefficient is mostly used similarity matrix to group similar dataset points in the cluster, which exactly identifies dissimilar cluster data points.

Clustering technique (GenClus) [17] designed for gene expression dataset samples, in addition it also deal with the problems of incremental gene expression dataset samples. This clustering method is developed based on the procedure of density based clustering approach. It doesn't depend on any clustering similarity measure to perform the clustering; it is easily applicable to larger dataset gene samples. This method exactly identifies the cluster data points, the clustering results is evaluated based on the z-score cluster validity quantify.

III.PROPOSED AFFINITY SEARCH CLUSTERING FOR CANCER

In this work we presents a novel clustering method based on the affinity search algorithm where in the proposed work the data of gene expression data is clustered based on similarity between two gene samples . In this step, the affinity search technique is referred from CAST method [18] in addition the proposed work the data is subdivide into several number of clusters for gene expression data , in earlier work CAST is applied for bioinformatics applications , since how number of clusters is not initially predefined by user. In the proposed the number of clusters for gene expression data is automatically decided based on the number of genes data in the dataset .So the proposed work will be more efficient than the other clustering methods .The similarity measure between the two cluster gene datasets samples is represented as $A_{n\times n}$for gene expression data samples D_i & D_j. Euclidean Distance (ED) measures is used as dissimilarity measure between the gene expression dataset samples .The distance similarity between two gene samples is mathematically specified as ,

$$A_{ij} = dis_{ED}(D_i, D_j) = \sqrt{\sum_{i=1}^{n} (D_i - D_j)^2} \quad (1)$$

In the above function square root function is removed automatically by introducing is monotonic and reverts to the equivalent position in clustering [19]. As formerly mentioned, subclusters for gene dataset samples are created successively through an affinity predefined threshold. Through significant the precise affinity predefined threshold rate, the cluster accept the high affinity gene expression dataset samples. The affinity threshold α is specified to decide which cluster gene dataset points are similar. The process of adding to and removing gene dataset samples to the subcluster is performed until all the gene dataset samples is completed. After each subcluster is constructed, a prototype is defined for each subcluster. The construction of a gene expression dataset samples follows the procedure in [20-21]. In the present work gene expression dataset samples are clustered based on the defined affinity threshold α for each gene dataset samples through the subcluster. Known the subclusterSC_i, its example is distinct through a gene expression dataset samples $R_i = \{r_1, \dots r_n\}$is then determined as,

$$r_x = \frac{\sum_{y\in SC_i} a_i(D_y) * d_{yx}}{|SC_i|} \quad (2)$$

where $D_y = \{d_{y1}, \dots . d_{yx}\}$ is a gene expression dataset samples and $|SC_i|$ designate the amount of gene expression dataset samples in the cluster. So the similarity among two gene dataset subclusters samples is determined and denoted as new similarity matrix $B_{M\times M}$ where B_{ij} is the defined as the similarity among two gene dataset for subclusters SC_i and SC_j. Dissimilarity among two gene dataset subclusters samples is determined based on the discrete wavelet (DTW) which is represented as similarity matrix Dis_{DTW} is dissimilarity among two gene dataset for subclusters SC_i and SC_j. Consider an example $R_x = \{r_{x1}, \dots r_{xn}\}$ then r_{x1} is determined from (2). To compute similarity between two newly modified clustered points SC_x and SC_y, is defined as $Z(R_x, R_y)$, where $Zi,j = dis_{ED}(r_{xi}, r_{yj})$ and disED( ) is the Euclidean distance. Given $W = \{w_1, w_2, \dots, w_u\}$ as a set of clustered data points from $w_u = \{(r_{x1}, r_{y1}), (r_{xi}, r_{yj}), \dots, (r_{xn}, r_{yn})\}$ is a set of points from Z , new distance $dis_{ED}(R_x, R_y)$ is determined by ,

$$dis_{ED}(R_x, R_y) = \min(\sum_{u=1}^{U} \frac{W_u}{U}) \quad (3)$$

Where $(r_{x1}, r_{y1}) = (1,1)$ & $(r_{xn}, r_{yn}) = (n, n)$ for $0 \leq r_{xi+1} - r_{xi} \leq 1$ and $0 \leq r_{yi} - r_{yi+1} \leq 1$ for all $i < n$, in this work the clustering of gene expression dataset samples follows the procedure of k-Medoids [22-23] ,Pseudo code is presented in algorithm 1.

A.Methods $(\boldsymbol{D}, \boldsymbol{\alpha}, \boldsymbol{K})$

Input : Gene dataset samples $D = (d_1, \ldots . d_n)$, α-affinty threshold ,k number of clusters for gene dataset samples

Output :C number of cluster for gene samples

(1) $A[N][N] \leftarrow Similarity_{ED}(D)$ from (2)
(2) $SC[1\ TOM], a\ [1\ to\ N]) \leftarrow CAST(A, \alpha)$/ *M is determined from CAST*/
(3) For (i=1 to M)
(4) $r \leftarrow Average(SC_i, a[1\ to\ N])$ / *summarize cluster*/
(5) $R \leftarrow R\ \cup r$
(6) End for
(7) Clustering
$B[M][M] \leftarrow Simialrity_{DTW}(R)$
(8) $C' \leftarrow k - mediods\ (k, B)$
(9) $C \leftarrow labels(C')$ validate the cluster results
(10) Return C

IV.EXPERIMENTATION RESULTS

In this section measure the performance accuracy results of the gencluster, hierarchical clustering and proposed Affinity search Clustering (ASC). In order to perform experimentation work we mainly focus on Cancer Genome Atlas (TGCA) [37], which is collected from National Human Genome Research Institutes (NHGRI) and National Cancer Institute's (NCI) and another types of the dataset is International Cancer Gene Consortium [34], (http://icgc.org/) .Both of these dataset is mainly used to mine the gene expression dataset samples for cancer prediction which consists of up to 50 different cancer types. In addition give outstanding suggestion designed for the scientists and researchers learning disease caner genetics and therapeutics. The clustering results are measured based on the following parameter

A.Precision (Pr): Precision is defined as percentages of predicted class which belongs to positive class that were correct, as determined using the equation:

$$\text{Precision} = \frac{A}{A+C} \quad (4)$$

B.Sensitivity (Sen): Sensitivity is defined as the percentage of predicted and actual class which belongs to positive cases that were correctly identified, as determined using the equation:

$$\text{Sensitivity(Sen)} = \frac{A}{A+B} \quad (5)$$

=(Number of true positive assessment)/(Number of all positive assessment)

C.Specificity (Spec): Specificity is defined as the percentage of predicted and actual class which belongs to negative cases that were correctly identified, as determined using the equation,

$$\text{Specificity (Spec)} = \frac{D}{D+C} \quad (6)$$

=(Number of true negative assessment)/(Number of all negative assessment)

D.Classification Accuracy (CA): Classification accuracy is defined as the percentage of the total amount of predictions which belongs to both positive and negative cases that were correctly identified, as determined using the equation:

$$\text{Classification accuracy (CA)} = \frac{A+D}{A+B+C+D} \quad (7)$$

=(Number of correct assessments)/Number of all assessments)

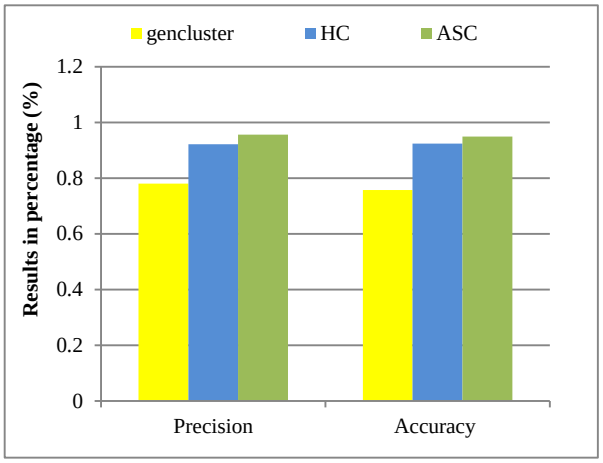

Figure 1. Precision and accuracy comparison vs. methods

The clustering accuracy results of proposed gencluster, hierarchical clustering and proposed Affinity search Clustering (ASC) is measured terms of the percentage of actual true positive results for Cancer Genome Atlas (TGCA) samples to identify cancer categories . An precision results of the ASC is 0.03% increased when compared to HC clustering and 0.1756% higher than the gencluster, so the test result shows that the contribution of the work is more accurate, regardless positive is illustrated in Figure.1. Similarly clustering accuracy results of ASC have achieved 0.9489 %, which is 0.0254% higher and 0.1921 % higher than HC and gencluster methods are illustrated Figure.1; these results are tabulated in Table 1

Parameters	Gencluster	HC	ASC
Precision	0.7804	0.9216	0.956
Accuracy	0.7568	0.9235	0.9489

Table 1. Precision and accuracy comparison for clustering methods

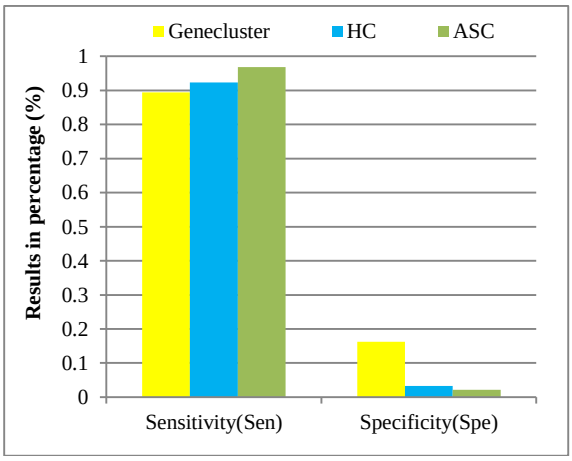

Figure 2.Sensitivity and Specificity comparison vs. methods

The sensitivity results of proposed gencluster, hierarchical clustering and proposed Affinity search Clustering (ASC) is represented as the percentage of actual true positive results for Cancer Genome Atlas (TGCA) samples to identify cancer categories. An precision results of the ASC is 0.0451 % increased when compared to HC clustering and 0.0746 % higher than the gencluster, so the test result shows that the contribution of the work is more accurate, regardless positive is illustrated in Figure.2. Similarly specificity results of ASC have achieved 0.0217%, which is 0.0108% lesser and 0.1406 % lesser than HC and gencluster methods is illustrated Figure.2.It shows that proposed ASC produces lesser error rate. These results are tabulated in Table 2.

Parameters	Genecluster	HC	ASC
Sensitivity(Sen)	0.8941	0.9236	0.9687
Specificity(Spe)	0.1623	0.0325	0.0217

Table 2. Sen and spe comparison for clustering methods

V. CONCLUSION AND FUTURE WORK

In this work presents a novel affinity search based clustering method for mining gene expression dataset samples in the biomedical applications which supports the recognition of helpful gene patterns based similarity measure. Specified the recent development of microarray technology, in this work we proposed a gene expression mining method for gene biomarkers. Mainly interested to mine the gene samples of the biomedical applications .Moreover, the mine gene expression data might be used for future work to recognize method of gene instruction and interface. The gene expression mining is performed based on the k mediods clustering method with affinity threshold value, experimentation results confirm that the proposed ASC produces higher results when compare to existing HC and gene cluster in terms of precision, sensitivity, accuracy and less for specificity .In the future work we apply the present work other image segmentation dataset samples and faulty genes are identified during clustering process for analysis of the diseases through less hoax.

REFERENCES

[1] M. Lan, Y. Xu, L. Li, F. Wang, Y. Zuo, Y. Chen, C.L. Tan, J. Su, "CpG-Discover: A Machine Learning Approach for CpG Islands Identification from Human DNA Sequence", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, pp 1702-1707, June 14-19 2009.
[2] Haque, S. Sahay, Y. Liu, "Investigation into Biomedical Literature Classification using Support Vector Machines", Proceedings of IEEE Conference on Computational Systems Bioinformatics, pp 366-374, 2005.
[3] Y.S. Pyon, J. Li, "Identifying Gene Signatures from Cancer Progression Data Using Ordinal Analysis", Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, pp 136-141, 2009.
[4] K. Amano, H. Nakamura, "Self-Organizing Clustering: "A Novel Non-Hierarchical Method for Clustering Large Amount of DNA Sequences", Journal of Genome Informatics", Vol. 14, pp 575-576, 2003.
[5] Siddiqui, G. Robertson, M. Bilenky, T. Astakhova, O.L. Griffith, M. Hassel, K. Lin, S. Montgomery, M. Oveisi, E. Pleasance, N. Robertson, M.C. Sleumer, K. Teague, R. Varhol, M. Zhang, S. Jones, "Cis-Regulatory Element Prediction in Mammalian Genomes", Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops, pp 203-206, 2005.
[6] S. Volinia, N. Mascellani, J. Marchesini, A. Veronese, E. Ormondroyd, H. Alder, J. Palatini, M. Negrini, C. M. Croce, "Genome wide Identification of Recessive Cancer Genes by Combinatorial Mutation Analysis", www.plosone.org, Vol. 3, Issue 10, pp 1-13, 2008.
[7] Data and Statistics. World Health Organization, Geneva, Switzerland, 2006.
[8] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics," CA: A Cancer J. Clinicians, vol. 64, pp. 9–29, 2014. "International Agency for Research on Cancer (IARC)," WHO, B. W. Stewart and C. P. Wild eds., World Cancer Report, 2014.
[9] A. K. Alqallaf, A. H. Tewfik, "Signal Processing techniques and statistics for the analysis of human genome associated with behavior abnormalities", proceedings of IEEE International conference on SSP, pp 36-38, 2007.
[10] D. Barash, D. Comaniciu, "Meanshift Clustering for DNA Microarray Analysis", Proceedings of the IEEE Conference on Computational Systems, pp 578-579, 2004.
[11] F.J. Lopez, M. Cuadros, A. Blanco, A. Concha, "Unveiling Fuzzy Associations between Breast Cancer Prognostic Factors and Gene Expression Data", Proceedings of 20th International Workshop on Database and Expert Systems Application, pp 338-342, 2009

[12] R. Xue, J. Li, D.J. Streveler, "Microarray Gene Expression Profile Data Mining Model for Clinical Cancer Research", Proceedings of the 37th Hawaii International Conference on System Sciences, pp 1-10, 2004.
[13] K. Raza and A. Mishra, "A novel anticlustering filtering algorithm for the prediction of genes as a drug target," Amer. J. Biomed. Eng., vol. 2, no. 5, pp. 206–211, 2012.
[14] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," IEEE Trans. Knowl. Data Eng., vol. 16, no. 11, pp. 1370–1386, Nov. 2004.
[15] E. Shay, (2003, Jan.). "Microarray cluster analysis and applications" [Online]. Available: http://www.science.co.il /enuka/ Essays/ Microarray-Review.pdf.
[16] M. B. Eisen, T. P. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Nat. Acad. Sci. USA, vol. 95, no. 25, pp. 14863–14868, Dec. 1998.
[17] Sarmah, Sauravjyoti, and Dhruba K. Bhattacharyya. "An effective technique for clustering incremental gene expression data." IJCSI (2010):
[18] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," Journal of Computational Biology, vol. 6, no. 3-4, pp. 281–297, 1999.
[19] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," DataMining and Knowledge Discovery, vol. 7, no. 4,pp. 349–371, 2003.
[20] V. Niennattrakul and C. Ratanamahatana, "Inaccuracies of shape averaging method using dynamic time warping for time series data," in Computational Science—ICCS 2007, pp. 513–52 , 2007.
[21] S. Aghabozorgi, M. R. Saybani, and T. Y. Wah, "Incremental clustering of time-series by fuzzy clustering," Journal of Information Science and Engineering, vol. 28, no. 4, pp. 671–688, 2012.
[22] X.Zhang, J. Liu,Y.Du, andT. Lv, "Anovel clusteringmethod on time series data," Expert Systems with Applications, vol. 38, no. 9, pp. 11891–11900, 2011.
[23] B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: experimental studies and comparative evaluation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09), pp. 312–319, 2009.
[24] The Cancer Genome Atlas. (2013). [Online]. Available: http:// cancergenome.nih.gov/
[25] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao,and A. Kasprzyk, "International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data," Database,vol. 2011, 2011

