Tanimoto Coefficient Similarity based Mean Shift Gentle Adaptive Boosted Clustering for Genomic Predictive Pattern Analytics

Marrynal S Eastaff, V Saravaan

Abstract: ***Gene expression data clustering is a significant problem to be resolved as it provides functional relationships of genes in a biological process. Finding co-expressed groups of genes is a challenging problem. To identify interesting patterns from the given gene expression data set, a Tanimoto Coefficient Similarity based Mean Shift Gentle Adaptive Boosted Clustering (TCS-MSGABC) Model is proposed. TCS-MSGABC model comprises two processes namely feature selection and clustering. In first process, Tanimoto Coefficient Similarity Measurement based Feature selection (TCSM-FS) is introduced to identify relevant gene features based on the similarity value for performing the genomic expression clustering. Tanimoto Coefficient Similarity Value ranges from ‘0’ to ‘1’ where ‘1’ is highest similarity. The gene feature with higher similarity value is taken to perform clustering process. After feature selection, Mean Shift Gentle Adaptive Boosted Clustering (MSGABC) algorithm is carried out in TCS-MSGABC model to cluster the similar gene expression data based on the selected features. The MSGABC algorithm is a boosting method for combining the many weak clustering results into one strong learner. By this way, the similar gene expression data are clustered with higher accuracy with minimal time. Experimental evaluation of TCS-MSGABC model is carried out on factors such as clustering accuracy, clustering time and error rate with respect to number of gene data. The experimental results show that the TCS-MSGABC model is able to increases the clustering accuracy and also minimizes clustering time of genomic predictive pattern analytics as compared to state-of-the-art works.***

Keywords : Genomic, Mean Shift Gentle Adaptive Boosted Clustering, Strong Learner, Tanimoto Coefficient Similarity, Weak Cluster, Weight

I. INTRODUCTION

A microarray database includes of many microarray gene expression data. The irrelevant features present in the microarray database increases time complexity of clustering algorithm. Higher dimensionality of the microarray database stimulates researchers to carry out feature selection using a variety of data mining techniques.

The conventional feature selection and clustering algorithms faced some issues such as large number of feature genes, fewer numbers of samples and lack of proper validation as gene expression data is prone to outliers and noise. In order to overcome such limitations, TCS-MSGABC model is developed in this research work using Tanimoto Coefficient Similarity Measurement based Feature selection (TCSM-FS) and Mean Shift Gentle Adaptive Boosted Clustering (MSGABC) algorithms.

Subspace Weighting Co-Clustering (SWCC) was performed in [1] for high dimensional gene expression data. However, clustering performance was not enhanced. Gene ontology (GO) annotations based semi-supervised clustering algorithm called GO fuzzy relational clustering (GO-FRC) was designed in [2] for clustering of microarray gene expression data. But, GO-FRC consumes more time for efficiently grouping the gene data.

A multivariate extension termed Relative Scan Statistics was carried out in [3] for comparison of two series in Bernoulli over frequent support. But, the clustering accuracy was not improved using Relative Scan Statistics. A random projection algorithm was introduced in [4] where random symmetric matrix was used to compute unsupervised clustering of dimensioned datasets like crystallographic structures. The clustering efficiency of PCA was not exact form of covariance/correlation matrix but it is symmetrical.

A new algorithm was introduced in [5] that measures similarity for individual gene groups and mixture of variants of hierarchical clustering to create the candidate groups. Semi-supervised consensus clustering (SSCC) was accomplished in [6] to enhance the robustness and quality of clustering results with a lower time complexity. However, error rate of gene expression data analysis was not reduced.

Rough-Fuzzy Clustering was carried out in [7] for grouping similar genes from microarray data with higher accuracy. But, clustering time taken for microarray data analysis was very higher. A novel gene selection method was developed in [8] depends on clustering where dissimilarity is estimated with help of kernel functions. However, clustering accuracy of gene expression data was poor.

Semi-supervised clustering was accomplished in [9] to resolve the gene expression data clustering problem with application of a multi-objective optimization. A feature selection based semi-supervised cluster ensemble framework (FS-SSCE) was employed in [10] for tumor clustering from bio-molecular data.

Revised Manuscript Received on November 30, 2019.

* Correspondence Author

Marrynal S Eastaff*, Ph.D Research Scholar, Department of Computer Science, Hinduthan College of Arts and Science, Coimbatore, India. Email: marrynalhindusthan@gmail.com

Dr. V. Saravanan, Associate Professor, Head Department of IT, Hindusthan College of Arts and Science, Coimbatore, India

To addresses the above mentioned existing issues in genomic predictive pattern analytics, TCS-MSGABC model is introduced. The main contribution of TCS-MSGABC model is depicted in below,

- To minimize the time complexity of genomic data pattern clustering when compared to state-of-the-art works, Tanimoto Coefficient Similarity Measurement based Feature selection (TCSM-FS) algorithm is introduced in TCS-MSGABC model. On the contrary to conventional works, Tanimoto coefficient is a popular similarity coefficients used to calculate the similarity between pairs of the gene features. Besides to that, Tanimoto coefficient is an association coefficient which is used for binary data, that assigned a value that range from 1 (represents the complete similarity) and 0 (denotes no similarity).
- To enhance the accuracy of genomic data pattern clustering as compared to conventional works, Mean Shift Gentle Adaptive Boosted Clustering (MSGABC) algorithm is proposed in TCS-MSGABC model. On the contrary to state-of-the-art works, MSGABC is AdaBoost algorithm which increases overall performance of clustering by reducing both the training error and generalization error than other existing works. Hence, MSGABC algorithm attains higher clustering accuracy for genomic predictive pattern analysis.

The remaining structure of the paper is created as follows. In Section 2, TCS-MSGABC model is explained with the assist of the architecture diagram. In Section 3, Experimental settings are presented and the experimental result of TCS-MSGABC model is discussed in Section 4. Section5 depicts the literature survey. Section 6 depicts the conclusion of the paper.

II. TANIMOTO COEFFICIENT SIMILARITY BASED MEAN SHIFT GENTLE ADAPTIVE BOOSTED CLUSTERING MODEL

The Tanimoto Coefficient Similarity based Mean Shift Gentle Adaptive Boosted Clustering (TCS- MSGABC) model is introduced in order to increases the performance of genomic predictive pattern analytics through clustering with a lower time. The TCS-MSGABC model is proposed by combining Tanimoto Coefficient Similarity measurement and Gentle Adaptive Boost Clustering algorithm on the contrary to conventional works. In proposed TCS-MSGABC, Tanimoto Coefficient Similarity measurement is a feature selection technique which selects a subset of relevant features for building robust genomic predictive pattern analytics models. Through removing most irrelevant and redundant features from input gene expression dataset, Tanimoto Coefficient Similarity measurement improves feature selection performance and thereby finds the important features with higher accuracy. After completing the feature selection process, TCS-MSGABC model utilizes the Gentle Adaptive Boost Clustering which is a variation of AdaBoost algorithm. The Gentle Adaptive Boost Clustering calculates one weak hypothesis during the each iteration and finally unites these weak hypotheses in a linear manner. The Gentle Adaptive Boost Clustering designs the strong clustering results by means of optimizing the weighted least square error in each run. In addition to that, the Gentle Adaptive Boost Clustering increases weights for wrongly clustered instances exponentially. Thus, Gentle AdaBoost exactly clusters the similar gene pattern data together with higher accuracy and a minimal amount of time complexity. Through an efficient clustering of gene data, proposed TCS-MSGABC model gets better genomic predictive pattern analytics performance. The architecture diagram of the TCS-MSGABC model is demonstrated in Figure 1.

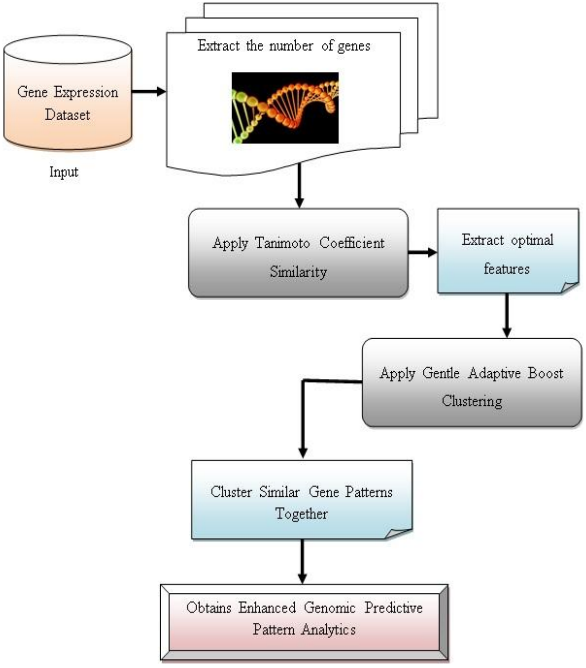
Gene Expression Dataset
Input
Extract the number of genes
Apply Tanimoto Coefficient Similarity
Extract optimal features
Apply Gentle Adaptive Boost Clustering
Cluster Similar Gene Patterns Together
Obtains Enhanced Genomic Predictive Pattern Analytics

Fig. 1. Architecture Diagram of TCS-MSGABC Model

Fig 1 presents the overall processes of the TCS-MSGABC model to obtain improved genomic predictive pattern analytics performance with minimum time complexity. As demonstrated in above Figure 1, TCS-MSGABC model initially takes gene expression dataset (i.e. gene expression cancer RNA-Seq Dataset) as input which contains number of gene data and features. Subsequently, the TCS-MSGABC model applies Tanimoto Coefficient Similarity measurement with aim of discovering the most relevant features from the input gene expression dataset. Finally, TCS-MSGABC model applies Gentle Adaptive Boost Clustering with objective of grouping similar kind of gene pattern together. From that, TCS-MSGABC model significantly identifies interesting gene patterns from the given gene expression data set for efficient genomic predictive pattern analytics. The exhaustive process of TCS-MSGABC model is explained in the subsequent subsections.

III. ALGORITHMS

A. Tanimoto Coefficient Similarity Measurement based Feature selection

Microarray gene expression data plays a significant role in feature selection as it supports for diagnosis and treatment of a variety of diseases.

Microarray gene expression data includes redundant genes feature in high dimensionality.

To reduce the time complexity of genomic predictive pattern analytics, a novel feature selection algorithm called Tanimoto Coefficient Similarity Measurement based Feature selection (TCSM-FS) is designed in TCS-MSGABC model. On the contrary to existing works, TCSM-FS algorithm is employed in proposed TCS-MSGABC model to find the significant gene features from an input Microarray gene expression dataset.

The TCSM-FS algorithm measures similarity between gene features. Based on measured similarity value, TCSM-FS algorithm remove irrelevant gene features that contain no useful information for the genomic pattern prediction and also remove redundant gene features that duplicate much or all of the information contained in one or more other features. From that, the TCSM-FS algorithm extracts the relevant gene features to reduce the time complexity of gene expression data analysis. Thus, TCSM-FS algorithm significantly avoids the curse of dimensionality for effective genomic predictive pattern analytics.

Let us assume an input microarray gene expression dataset contains number of gene features denoted as '$\tau_1, \tau_2, \tau_3, \ldots, \tau_M$'. Here, '$M$' denotes total number of gene features in a given dataset. Followed by, the tanimoto coefficient similarity between the gene features is calculated mathematically as,

$$\emptyset(\tau_1, \tau_2) = \frac{M * \sum \tau_1 \tau_2}{\sum {\tau_1}^2 + \sum {\tau_2}^2 - \sum \tau_1 \tau_2} \quad (1)$$

From the above mathematical expression (1), '$\emptyset(\tau_1, \tau_2)$' signifies a tanimoto similarity coefficient value, 'M' point outs the number of gene features in microarray gene expression dataset, ' τ_1, τ_2' indicates a two gene features in dataset. Here, '$\sum {\tau_1}^2$' designates a sum of squared score of the gene feature 'τ_1' and '$\sum {\tau_2}^2$' signifies a sum of squared score of the gene feature 'τ_2' whereas '$\sum \tau_1 \tau_2$' refers the sum of the product of the paired score of τ_1 and τ_2. The tanimoto similarity coefficient value gives the output results from 0 to +1. In TCSM-FS algorithm, '+1' point outs the high similarity between the gene features and '0' denotes the low similarity between the gene features. Thus, TCSM-FS algorithm chooses gene features with high tanimoto similarity coefficient value for increasing the genomic predictive pattern analytics performance.

The algorithmic processes of TCSM-FS is explained as follows,

```
Input: Microarray Gene Expression Dataset, Number of gene
features '$\tau_1, \tau_2, \tau_3, \ldots, \tau_M$'
Output: Select relevant gene features
Begin
Step 1:   For each input gene features
Step 2:   Compute tanimoto similarity coefficient between the
          two gene features '$\tau_1, \tau_2$'
Step 3:   If '$\emptyset(\tau_1, \tau_2) = +1$' then
Step 4:   Choose the features for genomic predictive pattern
          analytics
Step 5:   Else
Step 6:   Remove the gene features
Step 7:   End if
Step 8:   End For
End
```

Algorithm 1 Tanimoto Coefficient Similarity Measurement based Feature selection

Algorithm 1 depicts the step by step processes of TCSM-FS algorithm. As demonstrated in above algorithmic steps, at first TCSM-FS algorithm takes numbers of gene features from the microarray gene expression dataset as input. Subsequently the tanimoto similarity coefficient between the gene features are determined to identify the relevant and irrelevant gene features. If the tanimoto similarity coefficient value is '+1', then TCSM-FS algorithm extract gene features as more relevant for clustering genomic patterns. Otherwise, TCSM-FS algorithm eliminates gene features. With the chosen gene features, the similar types of genomic patterns are clustered which resulting in minimizing the time complexity.

2.2 Mean Shift Gentle Adaptive Boosted Clustering

In TCS-MSGABC model, Mean Shift Gentle Adaptive Boosted Clustering (MSGABC) algorithm is a machine learning ensemble technique. The MSGABC algorithm changes the weak mean shift clustering results into the strong cluster for obtaining better genomic predictive pattern analytics performance. The MSGABC algorithm is a variant of AdaBoost classifier. In TCS-MSGABC model, Mean Shift clustering is taken as weak learner which does not give the higher clustering accuracy for genomic patterns analysis. Hence, MSGABC algorithm is proposed in TCS-MSGABC model through combining the number of weak clustering results into a strong learner. The constructed strong learner accurately groups the same types of genomic patterns together with a minimal time complexity. The process involved in Mean Shift Gentle Adaptive Boosted Clustering is presented in below Fig 2.

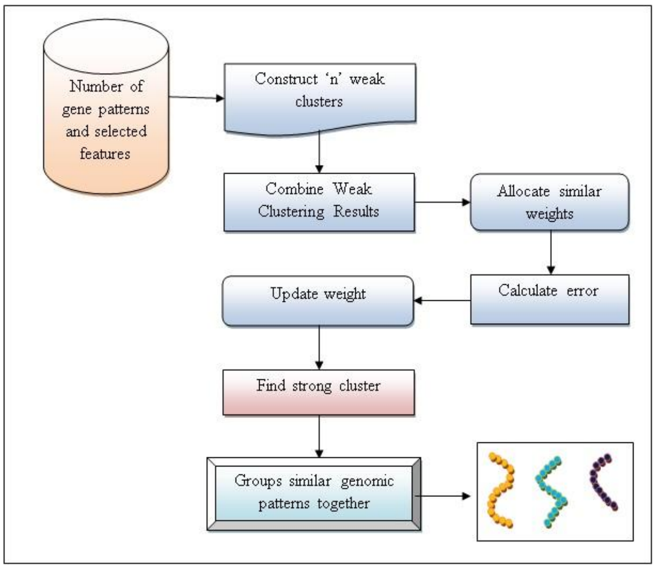
Number of gene patterns and selected features
Construct 'n' weak clusters
Combine Weak Clustering Results
Allocate similar weights
Calculate error
Update weight
Find strong cluster
Groups similar genomic patterns together

Fig. 2.Figure 2 Mean Shift Gentle Adaptive Boosted Clustering For Genomic Patterns Analysis

Fig 2 presents the block diagram of MSGABC algorithm for increasing the clustering accuracy of genomic patterns. As exposed in above figure, at first MSGABC algorithm designs a number of weak learner's i.e. Mean Shift Clustering for clustering the gene patterns. The weak learner is utilized in MSGABC algorithm is a centroid-based algorithm which works through grouping the each gene data to the mean of clusters in two-dimensional space. Thus, MSGABC algorithm removes irrelevant gene pattern data and thereby builds final set of clusters.

Let us consider a number of gene data in input microarray gene expression dataset is represented as '$\rho_1, \rho_2, \rho_3, \dots, \rho_m$' where '$m$' denotes number of gene data in a given dataset. In Mean Shift Clustering, the mean 'ϑ' is evaluated for all cluster in two dimensional spaces using below equation,

$$\vartheta = \frac{1}{m}\sum_{i=1}^{m}\rho_i \qquad \text{-- (2)}$$

From the above mathematical expression (2), 'ϑ' refers a mean of the cluster and 'ρ_i' point outs the gene data. Here, the mean is determined based on the weighted average of the gene data in two-dimensional space. Subsequently, the nearby gene data are grouped into the corresponding mean using below mathematical representation,

$$GKF\ (\vartheta, \rho_i) = \exp\left(-\frac{\|\rho_i - \vartheta\|^2}{2\ v^2}\right) \qquad \text{-- (3)}$$

From the above mathematical formulation (3), 'GKF' denotes a Gaussian kernel function whereas '$\|\rho_i - \vartheta\|^2$' represents a squared distance between the gene data and cluster mean in two dimensional space and 'v' signifies a deviation from its mean. During the every iteration, each input gene data is grouped into a nearest cluster mean. The process of Mean Shift Clustering is continual until all the gene data are grouped into the clusters. The clustering accuracy of weak learner is not sufficient for accurate genomic patterns predictions. As a result, the outputs of all weak clusters are combined into a strong learner using below,

$$\beta = \sum_{i=1}^{n} w_i(\rho) \qquad \text{-- (4)}$$

From the above formulation (4), 'β' indicates a strong clustering results. Here, $w_i(\rho)$ denotes the output of the weak cluster.

Followed by, MSGABC algorithm initializes a similar weight for each weak cluster. After that, MSGABC algorithm evaluates the error rate by considering squared differentiation between the actual and estimated output of the each weak learner using below,

$$\delta = \left(\beta_o - w_i(\rho)\right)^2 \qquad \text{-- (5)}$$

From the above expression (5), 'δ' indicates a training error of the each weak cluster whereas 'β_o' signify the actual output of the weak cluster and '$w_i(\rho)$' denote the obtained output of the weak cluster. After finding the error, the weight of the all weak cluster is updated as follow,

$$\omega\ (t+1) = \ \omega_{\ i}(t) * e^{-\beta_i w_i(\rho)} \qquad \text{-- (6)}$$

From the above mathematical representation (6), '$\omega\ (t+1)$' refers updated weight of the each weak cluster. Here, $\omega_i(t)$ denotes the initial weight of the weak cluster. Afterward the MSGABC algorithm identifies the weak cluster with minimum error as a strong learner. When the weak learner clusters the gene data incorrectly, the weight is increased. Otherwise the weight of the weak cluster is decreased in MSGABC algorithm. From that, the strong clustering results are obtained mathematically as,

$$\beta\ = arg\ min\ \delta * \ w_i(\rho) \qquad \text{-- (7)}$$

From the above formula (7), 'β' signifies a final output of the strong learner for clustering gene data with higher accuracy. Here, '$arg\ min\ \delta$' represents a minimum error of the weak cluster . With the support of obtained strong learner, MSGABC algorithm correctly groups all the input gene data patterns into the different clusters with a minimal error rate. The algorithmic process of the MSGABC is described in below,

Mean Shift Gentle Adaptive Boosted Clustering Algorithm

Input: Number Of Gene Data '$\rho_1, \rho_2, \rho_3, \dots, \rho_m$' and Number of Gene Features '$\tau_1, \tau_2, \tau_3, \dots, \tau_M$'

Output: Group similar gene data pattern together with higher accuracy

Begin

Step 1: For each gene data 'ρ_i'

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
Step 2:   Build 'n' weak clusters using (2) and (3)
Step 3:   Combine all weak clusters results using (4)
Step 4:   For each weak cluster '$w_i(\rho)$'
Step 5:       Define similar weights
Step 6:       Calculate error '$\delta$' using (5)
Step 7:       Update the weight '$\omega(t+1)$' using (6)
Step 8:       Discover strong learner using (7)
Step 9:   End for
Step 10:  Get strong clustering results for genomic patterns analysis
Step 11: End for
End
```

Algorithm 2 Mean Shift Gentle Adaptive Boosted Clustering

Algorithm 2 presents the step by step process of MSGABC. At first, MSGABC algorithm creates a number of weak clusters for each input gene data. Subsequently, all the weak clustering results are aggregated. Then, the similar weight is given for all the weak clusters. Subsequently MSGABC algorithm computes training error for each weak clustering result. Consequently, the weight of each weak cluster is updated based on determined error value. Finally, the MSGABC algorithm finds the weak cluster with a minimal error rate as strong learner. This strong learner precisely clusters the similar gene data pattern together with a lower time complexity. By an effective clustering of gene data, TCS-MSGABC model obtains better genomic predictive pattern analytics performance with the minimum false positive rate as compared to conventional works.

IV. EXPERIMENTAL SETTINGS

In order to validate the proposed performance, TCS-MSGABC model is implemented in Java Language using gene expression cancer RNA-Seq Dataset [21] from UCI machine learning repository with 20531 attributes and 807 instances. This dataset contains random extraction of gene expressions of patient's i.e. different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD. For conducting the experimental process, TCS-MSGABC model obtains various number of gene data in the range of 50-500 from gene expression cancer RNA-Seq Dataset. The efficiency of TCS-MSGABC model is evaluated in terms of clustering accuracy, clustering time and error rate with respect to diverse number of gene data.

The experimental evaluation of TCS-MSGABC model is conducted for several instances with respect to diverse number of gene data and averagely ten results are depicted in tabulation and graph. The experimental result of TCS-MSGABC model is compared against with two conventional works namely Gene ontology based fuzzy relational clustering (GO-FRC) [1] and Subspace Weighting Co-Clustering (SWCC) [2].

V. RESULTS

In this section, the comparative result of TCS-MSGABC model is presented. The performance of TCS-MSGABC model is compared against with Gene ontology based fuzzy relational clustering (GO-FRC) [1] and Subspace Weighting Co-Clustering (SWCC) [2] respectively. The effectiveness of TCS-MSGABC model is measured along with the following metrics with the help of tables and graphs.

A. Clustering Accuracy

Clustering accuracy 'CA' calculates the ratio of number of gene data correctly grouped to the total number of gene data taken for experimental process. The clustering accuracy is measured as follows,

$$CA = \frac{y_{ac}}{m} * 100 \qquad (8)$$

From the above mathematical equation (8), 'y_{ac}' signifies number of gene data accurately clustered and 'm' indicates a total number of gene data. The clustering accuracy is determined in terms of percentage (%).

Sample Mathematical Calculation:

- **Existing GO-FRC:** Number of gene data precisely clustered is 31 and the total number of gene data is 50. Then the clustering accuracy is obtained as,

$$CA = \frac{31}{50} * 100 = 62\ \%$$

- **Existing SWCC:** Number of gene data properly clustered is 36 and the total number of gene data is 50. Then the clustering accuracy is acquired as,

$$CA = \frac{36}{50} * 100 = 72\ \%$$

- **Proposed TCS-MSGABC:** Number of gene data perfectly clustered is 45 and the total number of gene data is 50. Then the clustering accuracy is determined as,

$$CA = \frac{45}{50} * 100 = 90\ \%$$

The tabulation result analysis of clustering accuracy for genomic pattern predictive analytics is depicted in below Table 1. When carried outing the experimental process using 400 gene data, the proposed TCS-MSGABC model achieves 93 % clustering accuracy whereas existing GO-FRC [1] and SWCC [2] obtains 78 % and 81 % respectively. Hence, it is considerable that the clustering accuracy of gene data patterns using proposed TCS-MSGABC model is higher when compared to other works [1] and [2].

Table I: Tabulation Result of Clustering Accuracy

Number of gene data (m)	Clustering Accuracy (%)		
	GO-FRC	*SWCC*	*TCS-MSGABC*
50	62	72	90
100	66	76	92
150	73	81	94
200	79	80	93
250	81	86	94
300	81	85	92
350	80	83	91
400	78	81	93
450	76	81	90
500	77	80	89

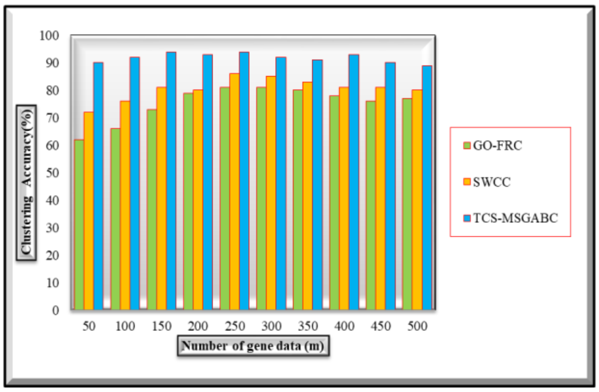

Fig. 3.Experimental Result of Clustering Accuracy versus Number of Gene Data

Fig. 3 presents the impact of clustering accuracy for genomic pattern predictive analytics with respect to various number of gene data using three methods namely GO-FRC [1] and SWCC [2] and proposed TCS-MSGABC model. As exposed in above graphical figure, proposed TCS-MSGABC model presents enhanced accuracy to group the similar genomic patterns together when compared with GO-FRC [1] and SWCC [2]. This is because of processes of TCSM-FS and MSGABC algorithms in proposed TCS-MSGABC model on the contrary to conventional works.

By using the concepts of the TCSM-FS algorithm, TCS-MSGABC model calculates tanimoto similarity coefficient between the gene features. Depends on estimated similarity value, TCSM-FS algorithm discovers the relevant and irrelevant gene features. After completing the feature selection process, TCS-MSGABC model utilizes MSGABC algorithm where it makes a strong learner with a minimal training error for efficiently cluster the gene data patterns based on selected features. Thus, proposed TCS-MSGABC model improves the ratio of number of gene data correctly grouped as compared to conventional works. Therefore, proposed TCS-MSGABC model obtains enhanced clustering accuracy for analyzing genomic patterns by 23 % as compared to GO-FRC [1] and 14 % as compared to SWCC [2] respectively.

B. Clustering Time

Clustering time calculates the time needed for clustering similar genomic data together. The clustering time is estimated as,

$$CT = m * t(csg) \qquad (9)$$

From the above formulation (9), '$t(csg)$' indicates a time utilized for grouping single gene data and 'm' signifies a total number of gene data. The clustering time is computed in terms of milliseconds (ms).

Sample mathematical calculation:

- **Existing GO-FRC**: time employed to cluster single gene data is 0.48 ms and the total number of gene data is 50. Then the clustering time is evaluated as,
  $$CT = 50 * 0.48 = 24\ ms$$
- **Existing SWCC:** the time consumed to cluster single gene data is 0.41 ms and the total number of gene data is 50. Then the clustering time is estimated as,
  $$CT = 50 * 0.41 = 21\ ms$$
- **Proposed TCS-MSGABC:** time required to cluster single gene data is 0. 36 ms and the total number of gene data is 50. Then the clustering time is computed as,
  $$CT = 50 * 0.36 = 18\ ms$$

The experimental result analysis of clustering time involved during process of genomic pattern predictive analytics is shown in below Table 2. When conducting the experimental evaluation by taking 300 gene data, the proposed TCS-MSGABC model gets 72 ms clustering time whereas conventional GO-FRC [1] and SWCC [2] attains 93 ms and 87 ms respectively. Therefore, it is significant that the clustering time of gene data patterns using proposed TCS-MSGABC model is lower as compared to other works [1] and [2].

Table II: Tabulation Result of Clustering Time

Number of gene data (n)	Clustering Time (ms)		
	GO-FRC	*SWCC*	*TCS-MSGABC*
50	**24**	**21**	**18**
100	**40**	**33**	**31**
150	**59**	**48**	**44**
200	**74**	**60**	**54**
250	**90**	**75**	**65**
300	**93**	**87**	**72**
350	**105**	**91**	**84**
400	**116**	**108**	**92**
450	**126**	**117**	**104**
500	**130**	**125**	**115**

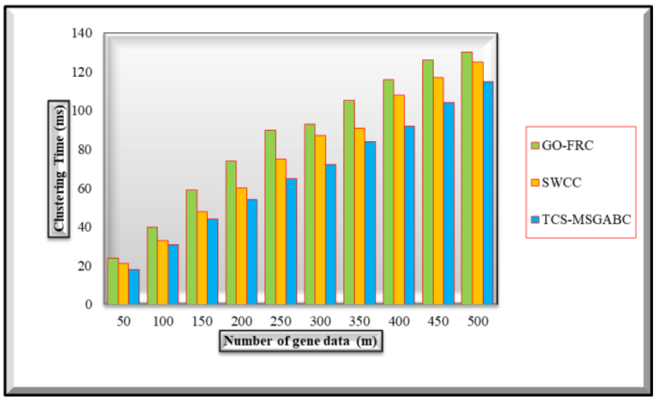

Fig. 4.Experimental Result of Clustering Time versus Number of Gene Data

Fig. 4 depicts the performance result of clustering time for genomic pattern predictive analytics based on different number of gene data using three methods namely GO-FRC [1] and SWCC [2] and proposed TCS-MSGABC model. As shown in above graphical diagram, proposed TCS-MSGABC model takes minimal amount of clustering time to group the related genomic patterns together when compared to GO-FRC [1] and SWCC [2]. This is owing to processes of TCSM-FS and MSGABC algorithms in proposed TCS-MSGABC model on the contrary to state-of-the-art works.

With the application of TCSM-FS algorithmic process, proposed TCS-MSGABC model take outs the significant gene features to decrease the time complexity of gene expression data clustering. From that, TCS-MSGABC model considerably minimizes the curse of dimensionality for effectual genomic predictive pattern analytics. Besides to that, proposed TCS-MSGABC model changes the weak mean shift clustering results into the strong cluster in order to correctly group the gene data with minimal time consumption. Accordingly, proposed TCS-MSGABC model minimizes the time required for clustering similar genomic data together as compared to conventional works. As a result, proposed TCS-MSGABC model attains reduced clustering time of genomic patterns analysis by 22 % as compared to GO-FRC [1] and 11 % as compared to SWCC [2] respectively.

C. Error Rate

Error Rate 'ER' computes the ratio of number of gene data wrongly grouped to the total number of gene data. The error rate is measured as,

$$ER = \frac{y_{ic}}{m} * 100 \quad (10)$$

From the above expression (10), 'y_{ic}' refers a number of gene data incorrectly clustered and 'm' denotes a total number of gene data in given dataset. The error rate is estimated in terms of percentage (%).

Sample Mathematical Calculation:

- **Existing GO-FRC:** number of gene data mistakenly clustered is 19 and the total number of gene data is 50. Then the error rate is obtained as,

$$ER = \frac{19}{50} * 100 = 38\ \%$$

- **Existing SWCC:** number of gene data poorly clustered is 14 and the total number of gene data is 50. Then the error rate is calculated as,

$$ER = \frac{14}{50} * 100 = 28\ \%$$

- **Proposed TCS-MSGABC:** number of gene data inaccurately clustered is 5 and the total number of gene data is 50. Then the error rate is measured as,

$$ER = \frac{5}{50} * 100 = 10\ \%$$

The performance result analysis of error rate involved during clustering process of gene data is demonstrated in below Table 3. When accomplishing the experimental work with 450 gene data, the proposed TCS-MSGABC model obtains 10 % error rate whereas state-of-the-art works GO-FRC [1] and SWCC [2] gains 24 % and 19 % respectively. As a result, it is expressive that the error rate of gene data patterns clustering using proposed TCS-MSGABC model is minimal when compared to other works [1] and [2].

Table III: Tabulation Result of Error Rate

Number of gene data (n)	Error Rate (%)		
	GO-FRC	*SWCC*	*TCS-MSGABC*
50	38	28	10
100	34	24	8
150	27	19	6
200	22	21	8
250	19	14	6
300	19	15	8
350	20	17	9
400	22	19	7
450	24	19	10
500	23	20	11

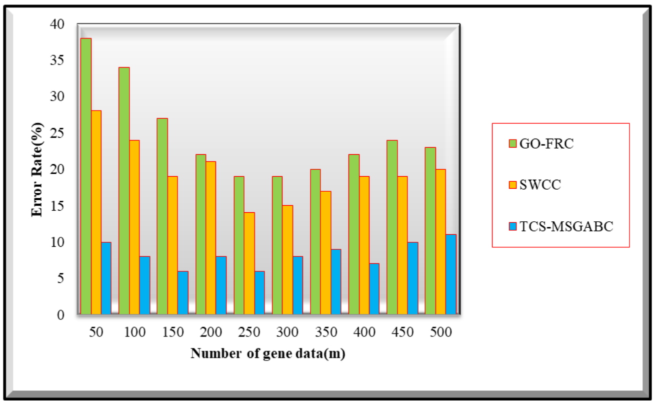

Fig. 5.Experimental Result of Error Rate versus Number of Gene Data

Figure 5 depicts comparative result of error rate for clustering genomic data patterns along with varied number of gene data using three methods namely GO-FRC [1] and SWCC [2] and proposed TCS-MSGABC model. As illustrated in above graphical illustration, proposed TCS-MSGABC model gives lower error rate to perfectly cluster the interrelated genomic patterns together when compared to GO-FRC [1] and SWCC [2]. This is due to processes of MSGABC algorithm in proposed TCS-MSGABC model on the contrary to state-of-the-art works.

By using the concepts of MSGABC algorithm, proposed TCS-MSGABC model get enhanced clustering performance for gene patterns analysis by considering both the training error and generalization error as compared to state-of-the-art works. As a result, proposed TCS-MSGABC model achieves better strong clustering result for grouping the genomic patterns. Hence, proposed TCS-MSGABC model reduces the ratio of number of gene data wrongly grouped as compared to conventional works. Consequently, proposed TCS-MSGABC model gets minimized error rate of gene data clustering by 65 % as compared to GO-FRC [1] and 57 % as compared to SWCC [2] respectively.

VI. LITERATURE SURVEY

Gradual shadowed set was utilized in [11] for grouping similar gene expression with a lower error rate. A Multiobjective Variable Length PSO-Based Approach was introduced in [12] for discovering non-redundant gene markers from microarray data and reducing time complexity.

Spectral ensemble biclustering (SEB) was employed in [13] for enhancing efficiency and scalability of gene expression data. Noise Resistant Generalized Parametric Validity Index of Clustering was presented in [14] for gene expression data analysis.

A novel method was designed in [15] for grouping of short time-course gene expression data with dissimilar replicates. Mutual Information-Based Supervised Attribute Clustering was performed in [16] for determining biologically considerable gene clusters with excellent predictive capability.

Projective clustering ensemble (PCE) was employed in [17] to get better quality of clustering gene expression data through dimensionality reduction. A review of different techniques designed for analysis of microarray data was presented in [18].

Tight clustering algorithm was employed in [19] to minimize time complexity of large microarray gene expression data. An evolutionary uncertain data-clustering algorithm was designed in [20] to determine the similarities among sets of gene expression clusters.

VII. CONCLUSION

An effective TCS-MSGABC model is designed with aim of enhancing the performance of genomic pattern predictive analytics via performing clustering with higher accuracy and minimal time. The aim of TCS-MSGABC model is obtained with the support of TCSM-FS and MSGABC algorithmic process on the contrary to conventional works. The proposed TCS-MSGABC model attains enhanced ratio of number of gene data that are accurately clustered by designing a strong learner with a lower error as compared to state-of-the-art works. Moreover, proposed TCS-MSGABC model gets minimal amount of time complexity to efficiently group the genomic data patterns as compared to existing works. Thus, proposed TCS-MSGABC model gives better accuracy, time and error rate for genomic pattern predictive analysis performance as compared to state-of-the-art works. The efficiency of TCS-MSGABC model is evaluated in terms of clustering accuracy, clustering time, and error rate and compared with state of the art works. The experimental result demonstrates that TCS-MSGABC model provides better performance with an enhancement of clustering accuracy and minimization of clustering time to find similar gene patterns when compared to state-of-the-art works.

REFERENCES

1. Xiaojun Chen, Joshua Z. Huang, Qingyao Wu, Min Yang, “Subspace Weighting Co-Clustering of Gene Expression Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 16, Issue 2, Pages 352 – 364, March-April 2019
2. Animesh Kumar Paul, Pintu Chandra Shill, “Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data”, Biosystems, Elsevier, Volume 163, Pages 1-10, January 2018
3. Chun-Pei Cheng, Kuo-Lun Lan, Wen-Chun Liu, Ting-Tsung Chang, Vincent S. Tseng, “DeF-GPU: Efficient and effective deletions finding in hepatitis B viral genomic DNA using GPU architecture”, Methods, Elsevier, Volume 111, Pages 56-63, December 2016
4. Luigi Leonardo Palese, “A random version of principal component analysis in data clustering”, Computational Biology and Chemistry, Elsevier, Volume 73, Pages 57-64, April 2018
5. Inti A. Pagnuco, Juan I. Pastore, Guillermo Abras, Marcel Brun and Virginia L. Ballarina “Analysis of genetic association using hierarchical clustering and cluster validation indices” Genomics, Elsevier, Volume 109, Issue 5-6, Pages 1-8, July 2017
6. Yunli Wang, Youlian Pan, “Semi-supervised consensus clustering for gene expression data analysis”, BioData Mining, Volume 7, Issue 7, Pages 1-13, December 2014
7. Pradipta Maji, Sushmita Paul, “Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 10, Issue 2, Pages 286 – 299, March-April 2013
8. Huihui Chen, Yusen Zhang, Ivan Gutman, “A kernel-based clustering method for gene selection with gene expression data”, Journal of Biomedical Informatics, Volume 62, Pages 12-20, August 2016
9. Abhay Kumar Alok, Sriparna SahaAsif Ekbal, “Semi-supervised clustering for gene-expression data in multiobjective optimization framework”, International Journal of Machine Learning and Cybernetics, Volume 8, Issue 2, Pages 421–439, April 2017
10. Zhiwen Yu, Hongsheng Chen, Jane You, Hau-San Wong, Jiming Liu, Le Li, Guoqiang, “Double Selection Based Semi-Supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 11, Issue 4, Pages 727 – 740, 2014
11. Ankita Bose, Kalyani Mali, “Gradual representation of shadowed set for clustering gene expression data”, Applied Soft Computing, Elsevier, Volume 83, Volume 105614, Pages 1-34, 2019
12. Anirban Mukhopadhyay, Monalisa Mandal, “Identifying Non-Redundant Gene Markers from Microarray Data: A Multiobjective Variable Length PSO-Based Approach”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 11, Issue 6, Pages 1170 – 1183, 2014
13. Lu Yin, Yongguo Liu, “Ensemble biclustering gene expression data based on the spectral clustering”, Neural Computing and Applications, Springer, Volume 30, Issue 8, Pages 2403–2416, October 2018
14. Rui Fa, Asoke K. Nandi, “Noise Resistant Generalized Parametric Validity Index of Clustering for Gene Expression Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 11 , Issue4 , Pages 741 – 752, 2014

15. Ozan Cinar, Ozlem Ilk, Cem Iyigun, “Clustering of short time-course gene expression data with dissimilar replicates”, Annals of Operations Research, Springer, Volume 263, Issue 1–2, Pages 405–428, April 2018
16. Pradipta Maji, “Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification”, IEEE Transactions on Knowledge and Data Engineering, Volume 24, Issue 1, Pages 127 – 140, 2012
17. Xianxue Yu, Guoxian Yu, Jun Wang, “Clustering cancer gene expression data by projective clustering ensemble”, PLoS ONE, Volume 12, Issue 2, Pages 1-21, 2017
18. Shweta Srivastava, Nikita Joshi, “Clustering Techniques Analysis for Microarray Data”, International Journal of Computer Science and Mobile Computing, Volume 3 Issue 5, Pages 359-364, May- 2014
19. Bikram Karmakar, Sarmistha Das, Sohom Bhattacharya, Rohan Sarkar & Indranil Mukhopadhyay, “Tight clustering for large datasets with an application to gene expression data”, Scientific Reports, Volume 9, Article number: 3053, Pages 1-12, 2019
20. Atakan Erdem, Taflan_Imre GUNDEM, “E-MFDBSCAN: an evolutionary clustering algorithm for gene expression time series”, Turkish Journal of Electrical Engineering & Computer Sciences, Volume 25, Pages 3443-3454, 2017
21. Gene expression cancer RNA-Seq Dataset: https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

AUTHORS PROFILE

Marrrynal S Eastaff, Assistant Professor, Department of Information Technology, Hindusthan College of Arts and Science, Coimbatore. She has completed her MSc Information Technology, MPhil and is now pursuing her Ph.D from Bharathiar University, Coimbatore. She has over 15 publications in international referred journals. Her area of specialization is Datamining

Dr. V. Saravanan is a Professor and Head of Department of Information Technology at Hindusthan College of Arts and Science, Coimbatore. He did his M.Sc in Computer Science at the Bharadhidasan University, Trichy and his MCA at the Periyar University, Salem. He started his teaching Profession at Thanthai Hans Roever College Perambalur, Trichy in 1999. Later in 2004 he joined Hindusthan College of Arts and Science, Coimbatore. His teaching areas are Software engineering and mobile computing as it is his main area of interest. He did his M.Phil and Ph.D at Manonmaniam Sundaranar university, Tirunelveli. His Ph.D was on Wireless Networking in video streaming. He has over 30 publications in internationals referred journals. He has presented research papers at conferences, published articles and papers in various journals. He is renowned key note address in both national and international conferences.