xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 1): pp.2123-2130
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.232

# DEVELOPMENT OF HYBRID GENETIC DISCRETIZATION GENOMIC MODEL USING CORRELATION-BASED CLUSTERING TECHNIQUES

**Dr. Vijay Arputharaj,** Faculty of Computer Science, School of Science and Information Tech, Skyline University Nigeria, Kano, Kano State, Nigeria.
https://www.sun.edu.ng/personnel/dr-ahmed-abba-phd
**Dr. Ahmed Abba Haruna,** Faculty of Computer Science, School of Science and Information Tech, Skyline University Nigeria, Kano, Kano State, Nigeria.
https://www.sun.edu.ng/personnel/dr-vijay-arputharaj-phd
**Ms. Jyoti Rajwar,** Faculty of Microbiology, School of Science and Information Tech, Skyline University Nigeria, Kano, Kano State, Nigeria.
https://www.sun.edu.ng/personnel/mrs-jyoti-rajwar

**Abstract-** In bio medical science, health disorders and their characteristics have a huge relationship with the gene expressions. The key elements used for diagnosis and prediction of health disorder. The data mining technique has a huge impact and application in human genetics and gene sequence data analysis. The huge size of data in the electronic format is considered as the big data. The storing, transferring and mining of genetic information within a big data are posed to be the current challenges in the process of huge data analysis. Classification of gene database is one of the most fundamental but yet a challenging problem that exists in the field of bio medical engineering and bioinformatics. There are a number of competent gene classification models which exist in current practice. These classification models in general have been used for natural language processing text classification, image recognitions, data prediction, reinforcement training etc. Some materials and methods required in this research are Association Rules, Clustering genes, Correlation Clusters in Gene Sequence, Cluster Editing in Gene sequences, Correlation Based Clustering, Logistic Regression. This technology was followed by Support Vector Machine Classification which reduced the execution time but still inquired high computational cost due to increased number of iterations. The salient features of the proposed technique Correlation Based Clustering Algorithm include fastest execution time, it has reduced cost considerably. It has improved the accuracy of result and reduced the number of rules.

***Keywords*: Gene Sequencing, Gene mining, Correlation clustering.**

## I. INTRODUCTION

In bio medical science, health disorders and their characteristics have a huge relationship with the gene expressions. The cell is a basic functional unit of the living organisms. Gene is a basic component of DNA which is located in the nucleus of the human cell. DNA is made up of one percent of protein coding called genes, the rest of DNA are non coded genes. These coded DNA genes called exons and non coded DNA genes called introns are the key elements used for diagnosis and prediction of health disorder. The structured gene classes contain both the labeled gene expressions as well as unlabeled data samples. The data mining technique has a huge impact and application in human genetics and gene sequence data analysis. There are different fields in human genetic science applied in data mining techniques. Currently, the numerous researches in governmental sectors, health care providers, educational industries and informational sciences have increased the quantity of information. The huge size of data in the electronic format is considered as the big data. The storing, transferring and mining of genetic information within a big data are posed to be the current challenges in the process of huge data analysis. In biomedical sector, the genetic disorders along with their distinctiveness are quite associated to the terms of genes and thus the identification of gene disorders by using gene data mining technique for finding the diseases is a vital part of it. This study deals with the proposed approach of CBC-MLRC this is a combination of supervised and unsupervised machine learning techniques for data analysis. The clustering is done by CBC whereas the classification is done by MNBC. The objective of this research is to identify various gene sequences in biomedical inherent learning. The domain and sub domain used here are data mining and classification respectively. The tool is developed by using Java technology. This research is useful for gene expression recognition by framing the association rules in accordance with the support measure and confidence

measure on the input dataset. It will extract and filter the required gene sequences based on CBC technique from the above input big data. The proposed technology correlation based clustering creates the gene clusters followed by drafting association rules which are applied on testing data to filter the required gene sequences. Finally, MLRC algorithm is applied as a classification algorithm [1] to identify the class labels of the test gene sequence in a big dataset.

**Motivation and Objectives**

The motivation to work on this project was identified from the unsupervised technology support vector machine classification for gene classification. This existing technology resulted in high computational cost due to increased number of iterations in input gene dataset. In classification method, the existing approach support vector machine classification[2] was identified as expensive and less accurate for gene classification[3] Prior to this technology a multi-purpose heuristic algorithm called as MOEDA[4] was in existence this was an improvement of existing UMD Algorithm[4]. It works based on two main rules. The first rule was defined as Higher and the Fewer Rule which was used to evaluate and categorize individual gene sequences. The second rule was defined as Forcibly Decreasing Rule that was used for introducing identified prospective individuals. The proposed technology correlation based clustering will overcome the above increased iteration issues. It is also cost effective compared to support vector machine classification method which is in existing practice.

### 1.1. *Objectives*

Genes sequence analysis is a method of subjecting DNA Sequences to a broad range of systematic method in order to know its character, configuration, nature and characteristics. CBC and the MNBC for gene sequence data analysis has some important objectives. It aims to organize a diseased diabetic genes from a vast stream of DNA gene sequence elements present in massive group of copious statistical data. This technique also attempts to approve, determine methods and tools for analyzing diseased diabetic gene sequence data. It also establishes methodologies for classifying DNA gene sequences. It helps to interpret the results accurately and meaningfully. Above all, the main objective is to overcome the limitations of the existing Support Vector Machine Classification technology which incurs high computational cost. The correlation based clustering will reduce the cost by reducing the number of iterations

**Problem Statement**

The problem definition can be classified under two main domains with reference to the above research.

### 1.2. *BigData Analysis*

Big Data analysis is described as an inspection method of the information in a contrived database. A variety of fields in human genetics were data mining technique, which has been applied to overcome the below mentioned issues.

- When discrepancies arise in an individual's gene sequence.
- Variances noticed in collection of codec algorithms.
- When there is a threat or database information security issues that arise.
- Genetic parental evolution issues on gene studies.
- Detection of differences in genetic parameters.

Big Data analysis is a method of identifying the progressions of determining pattern within immense statistics and data. Similarly DNA Gene sequence classifications also plays a vital role in defining the problem.

### 1.3. *DNA Gene Analysis*

Gene Sequence analysis also plays an important role in defining the problems attributed to this study.

- To expose the development and genetic variety of DNA gene sequences and organisms.
- To study the gene sequences, to identify the resemblance and infer if the genes are related.
- To identify the inherent character of the DNA gene sequence which include lively sites, gene structures, distribution of introns and exons etc.
- To identify DNA gene sequence variations like the point mutation, in order to get genetic marker.
- To identify hereditary disease prediction in gene sequences, for further parental treatment enhancements.

**Research Contributions**

This research could be classified into three phases based on studies carried out and research progression pattern. Phase-1 is concerned with Gene Sequence Analysis wherein analysis and classification of Basic Genetic Sequences (Introns and Exons) are done. Phase-2 of the studies deal with Medical Diagnosis - Disease Prediction which is associated to analysis and classification Protein Sequences that helps in disease prediction- mutation diabetics. Phase-3 is related to Parental Comparison during which analysis and classification of parental gene comparisons were carried out, this helps in forensic sciences.

In medical diagnosis, gene data mining techniques helps to identify various associations between the DNA genes based progressions and inconsistency in disease infections transformations. Currently, in bio technology gene sequence examination plays a major role in development of pharmaceutical therapies, recombinant gene techniques, applied immunology and diagnostic tests. In forensic biology, the gene sequence analysis is useful in anthropology studies, where as in humans, skeleton remains are used for determining race, gender, age group and other diseases. In the field of virology the gene sequences are used to identify the parasites components of the host cell that cause reproductive disorder, immunity problems etc. These are the various areas in which gene sequence analysis has its scope and applicability.

The research consists of an elaborated review of existing systems in practice for classifying DNA Gene Sequence Data Analysis. Initially a multi-purpose heuristic algorithm called MOEDA was in existence which was a development of UMD Algorithm. The main drawback of this system was high computational cost and high execution time. This technology was followed by Support Vector Machine Classification which reduced the execution time but still inquired high computational cost due to increased number of iterations. The salient features of the proposed technique Correlation Based Clustering Algorithm include fastest execution time, it has reduced cost considerably. It has improved the accuracy of result and reduced the number of rules.

Hung-Yi Lin (2016) has reviewed and knowledge on gene discretization and EM clustering was gained. The gene discretization depending on EM clustering and sequential forward gene selection for molecular classification gives us an in site about the character selection in huge quantizes of Big data. Bio informatics application with characteristic selection and dimensionality decrease technique for finding useful genes or choosing genes with discriminative influence has found to be extremely useful. Hence gene discretization depending on EM clustering has been employed to reduce complexities and enable better discrimination capability among identified genes. A sequential advance search algorithm has helped to explore the identified distinctive subset of genes containing discriminative influence. By studying the data obtained from selected characteristics, we can differentiate among numerous sub classes. It is also stated that experimental results have demonstrated the possibility of cancel classification depending on discretized gene appearance controlling[5].

Shruti Mishra et al(2016) Research on improved gene position approach using the customized trace ratio algorithm for gene appearance data was reviewed; it gives a view on how micro array technology enables understating information on gene characteristics by studying dimensional datasets. It has been stated that micro array characteristic data have been evaluated for fundamental biological mechanism of diseases, by building a gene regulatory network (GRN). One of the main prospects of the GRN process is gene selection considering a wide range of desirable gene sequence required for constructing the system. This can be done by two suitable methods as proposed in the existing research. The primary approach includes the gene assortment method called information gain, in which datasets are merged with other diverse algorithm called trace ratios. The other method is attributed to the execution of customized TR algorithm, to determine the weight age by scoring method. The efficiencies of both the process were evaluated in various classifier variants which include synthetic neural network classifier such as resilient propagations, rapid propagations, and reverse propagations and also SVM classifier. As a result of study it

has been observed the above proposed methodologies worked well with high accurateness and less iterations when compared to original TR algorithm[6].

## II. MATERIALS AND METHODS

Classification of gene database is one of the most fundamental but yet a challenging problem that exists in the field of bio medical engineering and bioinformatics. There are a number of competent gene classification models which exist in current practice. These classification models in general have been used for natural language processing text classification, image recognitions, data prediction, reinforcement training etc. Materials required for this research are as follows:

### *2.1. Big Data Analysis*

**Association Rule:** Association rule is defined as machine learning method for identifying interesting gene relationships between variables in huge gene databases.

It is proposed to recognize tough rules in gene databases using some method of relationships between the genes.

Support Rule - This rule is an indication to find out how frequently the coded gene proteins appear in the genome database. The supporting of X genes with respect to T is described as the proportion of diseased diabetic gene sequence t in the dataset which may have the gene protein X.

Confidence Rule - This rule is an indication to identify how often the relationship rules framed have found to be accurate. The assurance value of a rule X=>Y in context to a set of diseased diabetic gene sequence T is described as the ratio of diseased diabetic gene sequence which contains X which also contains Y.

The Lift Rule - The lift rule is described as the proportion of observed support to that which is likely if X and Y where considered independent.

Rule Power Factor - This rule is hint of how strong a rule's items are connected with each other in conditions of positive relationship.

Association Rule Application Method - Association rules are implied to convince a user-specified least support and a user-specified least confidence at the identical time. Association rule making is done using two separate steps:

- A minimum support verge needs to be applied to establish all frequent genes in a database.
- A minimum confidence limitation is applied to these repeated genes in order to form rules.

Sequence Pruning

Programmed DNA sequences are producing poor quality reads, especially near the sequencing introduction site, and the closing stages of extended gene sequence run. The genes from DNA library usually contain vector sequence more often. The poly A tails and other unrelated sequence are the common occurrences noticed in gene sequence analysis. Amplified exons are usually bordered by introns and primer sequences. If these gene sequences are not altered by trimming, any one of these artifacts will amend your progression assembly and downstream sequence study. Sequencer tends to provide easy-to-use but influential tools that help to alter and prune reduced quality and indefinite data.

- Trim Ends tends to remove deceptive data from the split ends of gene sequencing remains.
- Trim Vector tends to remove sequence-specific data altering the ends of the required gene sequences.
- Trim to Reference removes the edge of gene sequences that expand beyond an assembled orientation in a gene sequence.

Clustering genes: Clustering is defined as a process of grouping gene data points into clusters based on their correspondence.

Correlation Clusters in Gene Sequence: Establishes a process for grouping a set of genes into the best possible number of gene clusters without predicting the number of gene clusters[7] in advance.

Cluster Editing in Gene sequences: This functions in a context where the associations between the genes are known as a replacement for of the actual representations of the genes.

### *2.2. Correlation Based Clustering:*

A discrete optimization method is used to standardize the correlation based clustering functions. In this research work we proposed a probabilistic study of the gene sequence models that permits the correlation clustering function to estimate the original number of gene clusters. This analysis assumes that the function considers a standardized priority over all potential characteristics despite the number of clusters being formed. Clustering high-dimensional gene sequence data includes the analysis of gene data from a small number of clusters to numerous thousands based on scope. Correlation based clustering also attributes to diverse task, where correlations in the midst of different gene characters and characteristic vectors in space are assumed to survive guided by the association rules of the clustering process. These gene correlations tend to be diverse in different gene clusters. The global de-correlation will not decrease this to traditional (uncorrelated) clustering. Different spatial shapes of gene clusters being formed due to correlation amidst the subset of gene sequence data. Here, the connection between the various gene groups and their traits are described by considering the local correlation patterns. With reference to this understanding, the term correlation based clustering has been introduced concurrently with the perspective discussed above. Different stages of correlation based clustering are mentioned. Correlation based clustering is found to be closely associated to biclustering[8]. In case of biclustering the main purpose is to identify genes that share a correlation in some of their interrelated characteristics, where the correlation tends to be typical for every individual group.

Logistic Regression: It is a traditional statistic tool, which is associated to machine learning due to its association with SVM and Ad boost. They are commonly called as "large margin classifiers" since they attain the concept of margin either wholly or clearly. Large margin classifiers are generally framed by theoretical analysis as well as promising practices. Based on the values of the independent variables called predictors the logistic regression method is used to identify odd gene characteristics. The odds are described as the chance that a particular outcome is a gene feature divided by the chance that it is a non gene feature.

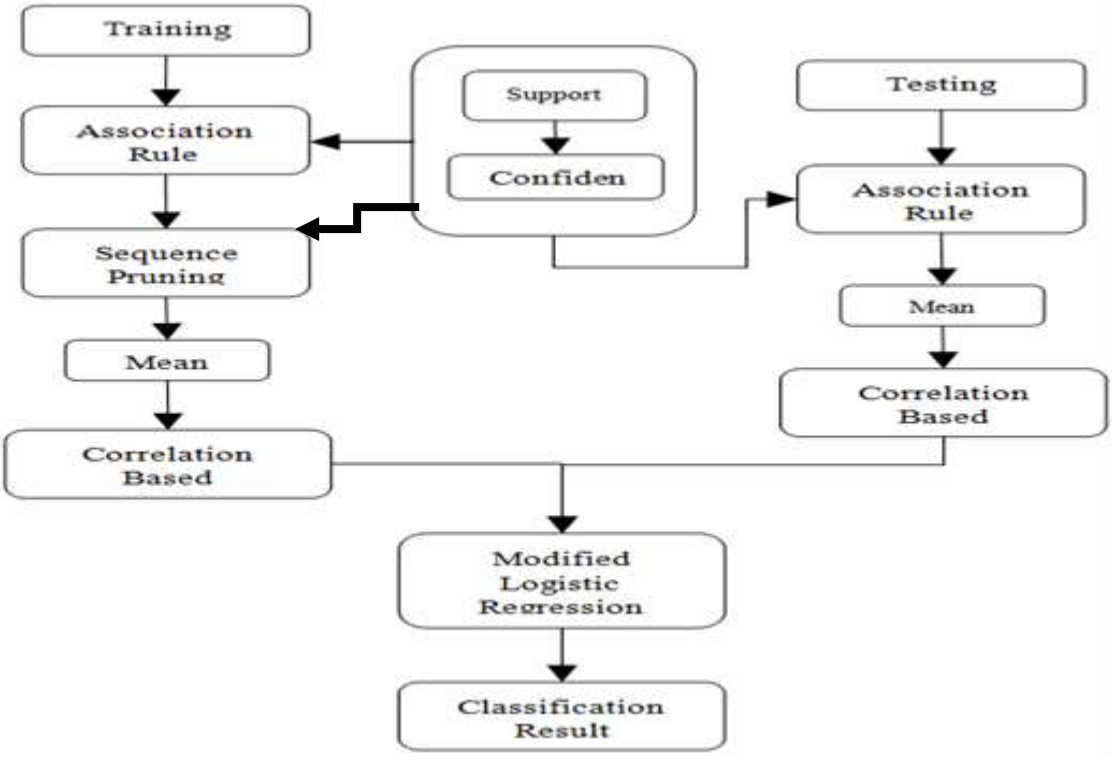


Figure 1. Flow of research

**Experimental Setup**

Different methods for correlation based clustering are available the relationship to different types of clusters are established using definite patterns[9]. This research over indenture during the evolution of

genetic algorithm with apposite protection surfaces in DNA genetic gene databases which was used Splice & HGMD Dataset.
The flow of this research contains 2 different datasets, the training dataset and testing datasets, after this process, the following step involves the development with the association rules applied, then the process of sequence pruning is implemented , mean finding, finally executing process of correlation-based clustering which is shown in detail in figure 1

- (CBC-MLGC) techniques are implemented to study the gene appearance detection process.
- Primarily, the training dataset contains different types of gene expression which in turn were used as input dataset to the structure that is to be accepted.
- The input dataset contained different gene sequence elements, illustration name and diverse group labels.
- The generations of related association rules were done with the aid of support and confidence rule, which has filtered the different gene sequences noticeably.
- The CBC technique was used to build the different clusters in the system environment.
- Then, the process of testing elements was initiated by providing testing dataset as input dataset to the system.
- Association rules were applied for assessing datasets with measurement of support rule and confidence rule calculation on the dataset. Then CBC was applied to the testing dataset.
- Finally MLRC was applied as classification algorithm to identify the group labels for the testing gene sequence dataset.

To authenticate the viability and presentation of the planned approach, executions are done in JAVA virtual machine. In this experimental process actual genetic material expressions data and artificial data were also used in Datasets.
Training Dataset: A training dataset consists of data's with examples which are used for knowledge, it is mainly used to study and fit the parameters. In most of the cases searching through training data for experimental relationships tend to over fit the data. This also means that they can identify and exploit noticeable relationships in the training data that do not exist in actual.
Testing Dataset: A test dataset is an independent dataset unlike the training dataset, but test data set also follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, it means minimal over fitting has taken place between the datasets. A better fitting of the training dataset as opposed to the test dataset usually points to over fitting of data[10].

The experimental process consists of 1000 examples selected at random from the absolute set of 3190 splice database[11]. Splice junctions are described as points on a DNA sequence where `superfluous' DNA is eliminated during the process of protein formation in higher organisms. The association rules are framed, within the dataset to distinguish and categorize the sequence of DNA at boundaries between exons the parts of the DNA sequence that are retained after splicing. Similarly introns the parts of the DNA sequence that are spliced out. This association rule framed consists of two main tasks. Firstly to identify exons/introns boundaries referred to as EI sites. Secondly identifying introns/exons boundaries referred as IE sites. In the biological synonyms, IE borders are compared to `acceptors' while EI borders are compared to donors.

## III. RESULTS AND DISCUSSION

The final results are tabulated as Clustering performance for splice dataset in table 1.1. The classification accuracy of proposed algorithm in tabulated in table 1.2. Comparison of multi class datasets to know classification accuracy is tabulated in table 1.3

**Table 1.1- Splice Dataset Clustering Evaluation**

| **Algorithm** | **Correctness** | **R O C** |
|---|---|---|
| Classifier 4.5 | 89.25% | 90.2% |
| Naïve Bayesian | 91.6% | 92.5% |
| SupportVectorMachine | 90.2% | 91.64% |
| simpleCart | 89.54% | 90.35% |
| KNN | 90.62% | 91.54% |
| CBC-MLRC | 92.87% | 93.12% |

Clustering performance has been replicated as diverse values of classifiers and outcomes attained were put into a table in Table 1.1. The significance of constraint correctness accuracy and ROC is taken for replication research studies. The actual code algorithm produces many of clusters repeatedly.

**Table 1.2- CBC-MLRC- accurateness for n genes**

| Genes | UFRF S | Alg1 | UFSFS | UFRDR | FRMIM | CFS | Proposed |
|---|---|---|---|---|---|---|---|
| 10 | 75 | 75 | 65 | 70 | 75 | 75 | 79 |
| 20 | 95 | 84 | 82 | 75 | 92 | 78 | 95 |
| 30 | 83 | 85 | 72 | 75 | 92 | 78 | 95 |
| 40 | 90 | 85 | 72 | 72 | 90 | 87 | 92 |
| 50 | 90 | 85 | 72 | 75 | 90 | 85 | 92 |

The actual simulation consequences mentioned in Table 1.2 depicts the projected classification accurateness in the proposed approach may present improved contrast to a variety of clustering algorithm when the number of cluster groups are very high.

**Table 1.3- CBC-MLRC Consolidated Accuracy with Multi Class**

| Data set | M O E D A | T S P | K-TSP | GA-ESP | KernelPLS+KNN | CBC-MLRC |
|---|---|---|---|---|---|---|
| Leukemiya | 99 | 97.1 | 97.1 | 96.5 | 99 | 99 |
| SRBCT | 95.6 | 95 | 99 | 98 | 96 | 98 |
| Lung | 95.7 | 83.6 | 94 | 90 | 95 | 97 |
| Splice Dataset | 96 | 96 | 95 | 95 | 95 | 96 |

The last table 1.3 depicts the quality of various multi class datasets used to get classification accuracy. The datasets used were listed in the first column; this is followed by various different algorithms used in the study. The efficiency of the proposed method is also proven in the given table. Proposed algorithms consume less execution time than all adaptive filters.
The above tables depict the effectiveness of the proposed algorithm, in terms of the performance measures which are stated above[12].

## IV. CONCLUSION AND FUTURE ENHANCEMENT

The Multi-Objective Heuristic Algorithm was proposed as a development of UMD algorithm which lacked in accuracy. This technology was followed by support vector machine classification which had high computational cost. Thus the proposed technique CBE- MLRC has overcome all the above mentioned limitations of the existing techniques. By using the data mining technique, the diversity of gene sequences has reduced considerably. The clustering technology has also helped to establish the sequences of extracted gene data. By comparing and filtering multi class gene cluster data, a determined accuracy has been attained in gene sequence dataset. The association rules which were drafted for the testing data with support and confidence calculation has found to be successful with reference to the above discussed results. The MLRC algorithm has also successfully supported the gene classification with above accurate results discussed. The execution time has also reduced considerable in this research.

As a future enhancement, this technology may be applied for the studies on larger scale databases in various challenging fields as mentioned below:

- This may also be applied in disease prediction and advanced research studies.
- It may be applied to researches in the field of chemical engineering. Sequence analysis composes of techniques that are used to find the sequence of polymer formed by several monomers[13]. This is compared to DNA sequencing in genetics and molecular biology.
- It can be applied to the field of marketing where the sequence analyzing techniques applied to study and manage analytical customer relationship applications such as NPTB models[14] (Next Product to Buy).

- In sociology sequence methods are mostly used to study and interpret life-course, career trajectories, patterns of establishment and national development etc. This body of research has further established rising subfields of social sequence analysis[15].

**Acknowledgments**

I would also like to thank my supervisors Dr.Sheeja S, Professor, KAHE, Dr.G.MANICKA CHEZIAN, Associate Professor, NGM College, for his valuable advices and suggestions that helped me in completing this research work. I would express my sincere thanks to Dr.D.VENI Ph.D., Head of Department of Computer Science, Karpagam Academy of Higher Education, for his timely efforts and advice in my research work.

## REFERENCES

[1] Sossi Alaoui, Safae & Farhaoui, Yousef & Aksasse, B.. (2018). Classification algorithms in Data Mining. International Journal of Tomography and Simulation. 31. 34-44.
[2] Eads, Damian & Hill, Daniel & Davis, Sean & Perkins, Simon & Ma, Junshui & Porter, Reid & Theiler, James. (2002). Genetic Algorithms and Support Vector Machines for Time Series Classification. Proc. SPIE. 4787. 10.1117/12.453526.
[3] Hori, Gen & Inoue, Masato & Nishimura, Shin-Ichi & Nakahara, Hiroyuki. (2002). Blind Gene Classification - An ICA-based Gene Classification/Clustering Method.
[4] Lv, Jia & Peng, Qinke & Chen, Xiao & Sun, Zhi. (2016). A multi-objective heuristic algorithm for gene expression microarray data classification. Expert Systems with Applications. 59. 13-19. 10.1016/j.eswa.2016.04.020.
[5] Hung-Yi Lin, Gene discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification, Applied Soft Computing, Volume 4, Issue 8 (2016) PP 683–690, 2016.
[6] Shruti Mishra, Debahuti Mishra, Enhanced gene ranking approaches using modified trace ratio algorithm for gene expression data, Informatics in Medicine Unlocked, Volume (5): Issue (1), PP 39-51, 2016
[7] Samal, Mamata & Saradhi, V. & Nandi, Sukumar. (2018). Scalability of correlation clustering. Pattern Analysis and Applications. 21. 10.1007/s10044-017-0598-7.
[8] Qin, Ruxin & Tian, Yingjie & Chen, Jing & Deng, Naiyang & Zhang, Haibin. (2009). Data mining method of association rule for bi-cluster. Journal of Beijing University of Technology. 35.
[9] Vijay Arputharaj J, Dr.Sheeja S, "An Analysis of Modified Naïve Bayesian Classification using Correlation based Clustering for Gene Sequence Data Analysis", International Journal of Engineering & Technology(UAE), Volume 7, Issue 4.5, Aug-Sep 2018, PP 612-616
[10] Vijay Arputharaj J, Dr.Sheeja S , "Correlation-based Clustering and the Modified Naïve-Bayesian-Classification for Gene-sequence data analysis" , International Journal of Engineering & Technology(UAE), Volume 7, Issue 4, Aug-Sep 2018, PP 5292-5299
[11] The SPLICE dataset: Classification - https://jmlr.csail.mit.edu/papers/volume1/meila00a/html/node32.html
[12] Dr.Vijay Arputharaj et.al, Basic Gene Discretization-Model using Correlation Clustering for Distributed DNA Databases, Int. J. Advanced Networking and Applications, Volume: 11 Issue: 05 Pages: 4407-4417(2020) ISSN: 0975-0290
[13] Polymerization - chemical reaction - https://www.britannica.com/science/polymerization
[14] Aaron Knott, Andrew Hayes, Scott A. Neslin,, Next-product-to-buy models for cross-selling applications, Journal of Interactive Marketing, Volume 16, Issue 3, 2002, Pages 59-75, ISSN 1094-9968,
[15] Sequence Data Mining - https://link.springer.com/content/pdf/bfm%3A978-0-387-69937-0%2F1.pdf