CORRELATION BASED CLUSTERING AND THE MODIFIED NAÏVE BAYESIAN CLASSIFICATION FOR GENE SEQUENCE DATA ANALYSIS

Vijay Arputharaj J
Research Scholar, Karpagam Unversity and Lecturer, Jigjiga University
phdvij@gmail.com

Dr.Sheeja
Head, Department of Information Technology, Karpagam University, Coimbatore
sheejaajize@gmail.com

Abstract Correlation based Clustering separates the statistical data into the most favourable amount of clusters on the correspondence to the statistically analysed data points. As we know that, Data mining is the technique of figuring out the progression of determines patterns inside huge statistics and data sets which concerning on techniques on the connection with machine related learning, statistics and also the advanced database systems.this technique denotes the gene sequence by using the novel classification technique, which improves the accuracy of classification under the curse of dimensionality, also clustering the gene data based on correlation based clustering will reduce the execution time.

***Keywords*:** Clustering; Classification, Data mining, Progression.

1. Introduction

The importance in Data mining technique in the process of human genetics are variety of applications are used, an significant objective exist to appreciate the mapped affiliation flanked by the personage disparity in human gene DNA genomic sequences and unpredictability in various code algorithms for information database sanctuary issues, for transformation vulnerability and parental comparisons, identification differences. In the country like India which is steadily occupied in attendance is gigantic requirement for DNA genetic databases which may facilitate in prevention in diverse forms of swindles as like Passport, License, Ration Card deceit, other supplementary hoax etc. Also quite a lot of prophecy or visualization and very advanced mining performances are available to enhance it, and these are used to authorize, effort to determine original process techniques for discriminate DNA gene sequences or exons, from non-coding DNA gene progression sequence or introns. So the information mining is the unsurpassed method to evaluate and extort the genomic data, this technique is moreover supportive to make the frequent code algorithm The very imperative distinctive point in the dataset is the information detection from the massive group of copious statistical data. Which move ahead in computer knowledge, in meticulous the network, enclose led to “data and sequence explosion”.

Nowadays, the numerous developments in the government, health care, education, science and information technology raises the density of information. The processing of large size data in electronic form regarded as the bigdata. The storage, transfer and the extraction of meaningful information from large scale data are the major processes in the Big Data analysis. In medical field, the diseases and their characteristics are related with the gene expressions and hence the recognition of diseases for diagnosis is the major task. The collection of large amount of labeled gene expression and the utilization of few unlabeled data samples govern the identification of structure of gene classes.

2. Literature Study

This entire chapter provides overview of research carried out on clustering algorithm and their application to several microarray data reported in literature. This chapter is broadly divided into two sections. First section deals with research work carried out on several microarray data and section two deals with research carried out on different clustering algorithm

Review work carried out on clustering data: DNA microarrays are high-throughput methods for analyzing complex nucleic acid samples. It makes possible to measure rapidly, efficiently and accurately the levels of expression of all genes present in a biological sample. The application of such methods in diverse experimental conditions generates lots of data. However, the main problem with these data occurs while analyzing it. Derivation of meaningful biological information from raw microarray data is impeded by the complexity and

vastness of the data. To overcome the problem associated with gene expression microarray data many statistical methods has been proposed in recent past. Some important has been explained below:

•ENFSI DNA Working Group April 2012: Studies on the statistics, performance and different search strategies of DNA databases are usually done using simulated DNA-databases some scientists however have asked for disclosure of the real DNA-profiles contained in DNA-databases to allow them to evaluate some of the population genetic assumptions underlying DNA-testing38 Of course this should be done under strict conditions and removing any links to the identity of the owner of the DNA-profile. Some countries do already allow this in the interest of quality assurance and/or process improvement. A big problem for DNA-database managers is that they cannot distinguish matches with monozygotic twins. Promising epigenetic research is going on42 but the amounts of DNA which are necessary for a test need to go down to be able to analyze forensic samples containing low amounts of DNA[1]

•Marina Andrade & Manuel Alberto M. Ferreira (2010): The use of DNA profiles in forensic identification problems has become, in the last years, an almost regular procedure in many and different situations. Among those are: 1)disputed paternity problems[2], in which it is necessary to determine if the putative father of a child is or is not the true father; 2) criminal cases as if a certain individual Y was the origin of a stain found in the scene of a crime; or in more complex cases to determine if an individual or more did contribute to a mixture trace found[2]; 3) civil identification problems[2], i.e., the case of a body identification, together with the information of a missing person belonging to a known family, or the identification of more than one body resultant of a disaster or an attempt. And even immigration cases in which it is important to establish family relations. To connect an individual with a crime on the basis of a profile match may be dangerous because the database may contain undetected errors. In order to avoid misclassification with DNA from the database it is important to admit, at least, a second and independent analysis. After computing the likelihood, whether it is a criminal case or a civil identification case, it is possible to compute the posterior odds, i.e., multiplying the likelihood ratio and the prior odds, in order to perform a comparative evaluation between the prosecution and the defense hypotheses. The database file α is a subset of the population set P,$\alpha \subset P$. If the size of the database file is small, then one may only have a small fraction of the possible offenders. Therefore, it is important to take that into account [2].

•V.N. Rajavarman and S.P. Rajagopalan (2007), the k-means algorithm helps us to discover associations genes-genes and genes-environmental factors. We have experimented the classical k-means algorithm without any feature selection. The execution time was very large (over 7500 minutes) and results cannot be interpreted (we didn't know which were the features involved in the disease) so the feature selection phase is required. With the feature selection, the time of execution of k-means had decreased to 1 minute and the results are exploitable. We present here clusters obtained with k = 2 and their number of occurrences. This shows that the k-means algorithm using results of the GA, is able to construct clusters very closely related to the solution presented in results of the workshop. Moreover this solution has been exactly found 4 times over 10 of executions[3].

The genetic algorithm managed to select interesting features and the k-means algorithm was able to class pairs of individuals according to these features and to confirm interesting associations of features [3].

•Chan Wai Keung Brian (2006): The advantage of using genetic algorithm[4] is that it doesn't have to know any rules of the problem in advance – the rule will can be found through evolution. This is very useful for very complex and loosely defined problem. The drawback of genetic algorithm is that the definition of the fitness function can be very complicated sometime. The fitness function may affect the performance of the process significantly if the complexity of the fitness function increase. It is because the fitness function is used to compare every element in the sample population to every data in the training data set. Sometimes an acceptable solution cannot be derived even after countless iteration if the genetic operators are wrong chosen.

•Shipp M. A., Ross K. N., Tamayo P., Weng A.P., Kutok J. L., Aguiar R. C., Gaasenbeek M., Angelo M., Reich M., Pinkus G. S., Ray T. S., Koval M. A., Last K. W., Norton A., Lister T. A., Mesirov J.,Neuberg D. S., Lander E. S., Aster J. C., Golub T. R. (2002): Diffuse large B-cell lymphoma (DLBCL), the most common lymphoid malignancy in adults, is curable in less than 50 % of patients. Prognostic models based on pre-treatment characteristics, such as the International Prognostic Index (IPI), are currently used to predict outcome in DLBCL. However, clinical outcome models identify neither the molecular basis of clinical heterogeneity, nor specific therapeutic targets. In this paper analysis has been done on the expression of 6,817 genes in diagnostic tumor specimens from DLBCL patients who received cyclophosphamide, adriamycin, vincristine and prednisone (CHOP)-based chemotherapy, and applied a supervised learning prediction method to identify cured versus fatal or refractory disease. The algorithm classified two categories of patients with very different five-year overall survival rates (70% versus 12%).

3. Related Work and Study

Gene is a region of DNA. It is used in the medical field to find the disease prediction. Small amount of genes are used for disease prediction means, it have less cost. If amount of genes increased means, it will improve the cost. Here a multi-objective heuristic algorithm called MOEDA is proposed. It is an improvement of Univariate Marginal Distribution Algorithm. There is two main rules are used. They are 1) Higher and Fewer Rule, 2) Forcibly Decrease Rule. First rule is used for evaluating and sorting individuals. Second rule is used to generate potential individuals. After that Support Vector Machine (SVM) classification is used for gene classification. The main drawback in this methodology is high computation cost. Cost is increased because of number of iterations.

To overcome the limitations above, we propose "Correlation Based Clustering and Modified Logistic Regression Classification" (CBC-MLGC) as the research for gene expression recognition. Initially, the training dataset containing various gene expressions are regarded as the input to the system. The input contains gene sequence, instance name and class labels. The generation of Association rules with the support and confidence measures filter the gene sequences considerably. Then, the Correlation Based Clustering (CBC) is used to create the clusters. Then, the testing is started by giving testing dataset as input. Association rules are used for testing data with support and confidence calculation. Then correlation based clustering is applied to the testing data. And finally modified logistic regression classification is used as classification algorithm to find the class labels for the testing gene sequence dataset.

4. Materials and Methods

Data mining and information retrieval: Data mining is the examination process of the Information invention in the DNA Genetic Databases, It is also an inter disciplinary broad area of computer field, it is the computation and calculation development of determine prototype in huge information data sets linking techniques at the connection of reproduction intellect, mechanism knowledge, information, and database schemes. The general objective of the information mining procedure is to haul out information in sequence from a large dataset and renovate it keen on an explicable constitution for auxiliary purpose. Away from the unprocessed examination pace, this process engross catalogs and data administration characteristics, Dataset statistics and preprocessing, DNA data model and Database supposition reflection, Database model metrics, intricacy concern, DNA Database post-processing of revealed constitutions, apparitions.

In the main, data mining is the progression of scrutinize data and statistics commencing dissimilar viewpoint and abbreviation it keen on constructive statistical DNA database information - in sequence it preserve to be old to enlarge returns, overheads, or equally. Data mining[5] is solo number of methodical utensils for considering statistics. Which also tolerate data abusers to study data statistics as of countless dissimilar magnitudes, sub group it, and which used to sum up the associations identified in DNA dataset operations? Precisely, mining technique is the progression of decision association or prototype in the middle of several numbers of turfs in huge relational DNA databases.

Even though data mining technique is a reasonably innovative tenure for the biological datasets as like DNA databases, the knowledge is not expertise in the particular sub field. Still incessant modernism in computer dispensation supremacy, computer disk storage space, and algebraic calculation are spectacularly growing the accurateness of examination whereas powerful losing in the rate.

For an illustration, one DNA datasets used the data mining facility of Oracle backend to evaluate confined sequence models. It exposed to facilitate while datasets diapers on specific values, they also tend to analysis showed that these datasets characteristically process the data items in the database items on it. The following are the specific aspects of data mining techniques are used.

Data, Information, and Knowledge

Data: Data in the DNA genetic places are any particulars, figures, or text that can be processed by a processor. at present, company groups are build up huge and rising quantity of data in dissimilar arrangement and dissimilar database sets. This also includes an operational or transactional data like raw values in dataset, nonoperational data, such as statistical data, and comprehensive informatics data

Meta data - data concerning data itself, such as rational DNA database plan or information statistical dictionary descriptions

Information-The prototypes, dealings, or relationships amongst everyone this information can give information about the datasets. For instance, psychoanalysis of DNA values in point of comparisons of parental data can surrender in turn on the processes on it.

Knowledge-Information and statistics can be rehabilitated into knowledge information about past patterns and future trends. For example, summary information on datasets can be analyzed in light of elements to supply knowledge of checking elements. Thus, an end-user could determine which data are most vulnerable to process constraints.

Data Warehouses- theatrical progress in DNA dataset and data incarcerate, dispensation supremacy, data information broadcast, and storeroom competences are facilitates association to put together their various databases keen on data information warehouses in process elements. Data warehousing processes are definite as a course of national information statistical administration and repossession.

Classes: The accumulated data information is used to place statistics in prearranged groups. For example, a mutation growth comparison classes could mine data to determine when they typically arrange the particular datasets in it. This information might be uses to augment dataset interchange by multiple classes[6].

Clusters: Data substances are clustered or intergroup accord to reasonable associations or end user preferences. [7]

Associations: Data can be excavated to make out relative associations. The main example is an example of associative mining.[8]

Sequential patterns: Data is excavated to await performance behavioral prototypes and inclinations.

Special stages of examinations are also available in data mining:

Artificial neural networks: Non-linear prognostic replicas that study from side to side preparation and are like genetic neural networks in arrangement.

Genetic algorithms: several optimization methods that use procedure such as hereditary mixture, mutation alteration, and expected collection in a plan based on the perception of natural DNA database development.[9][10]

Decision trees: Tree formed configurations that signify set of decisions. These verdicts cause policy and regulations for the arrangement of a dataset.

After the Data mining technique to knowledge discovery from entire DNA Database, There can be three levels of genome data mining. The simplest is an in-depth analysis of the result from a single query using a genome browser. In this level, one may start with a gene or marker name, or by mapping a sequence to the genome. Cross comparison of various annotation 'tracks' may help make sense of the query region. This is the most popular use of any genome browser[11]. Data mining is opposite to the information retrieval in the sense, it does not based on predetermine criteria; it will uncover some hidden patterns by exploring our data[12].

This proposed approach is a combination of protected refuge techniques explicitly digital substantiation and validation process followed with the data information mining in the DNA Genomic database. This research above all contract through the progression of genetic gene based algorithm with suitable safety textures in DNA Genomic Databases- Splice Dataset[13][14][15]

5. Performance Parameters

We analyze the performance of proposed work for number of rules, precision, recall, accuracy, execution time etc.

6. Flow of Research

The flow of research associated with the training and testing process of datasets, the next step involves with association rules, which followed by sequence pruning, mean, correlation based clustering which shown in figure 1

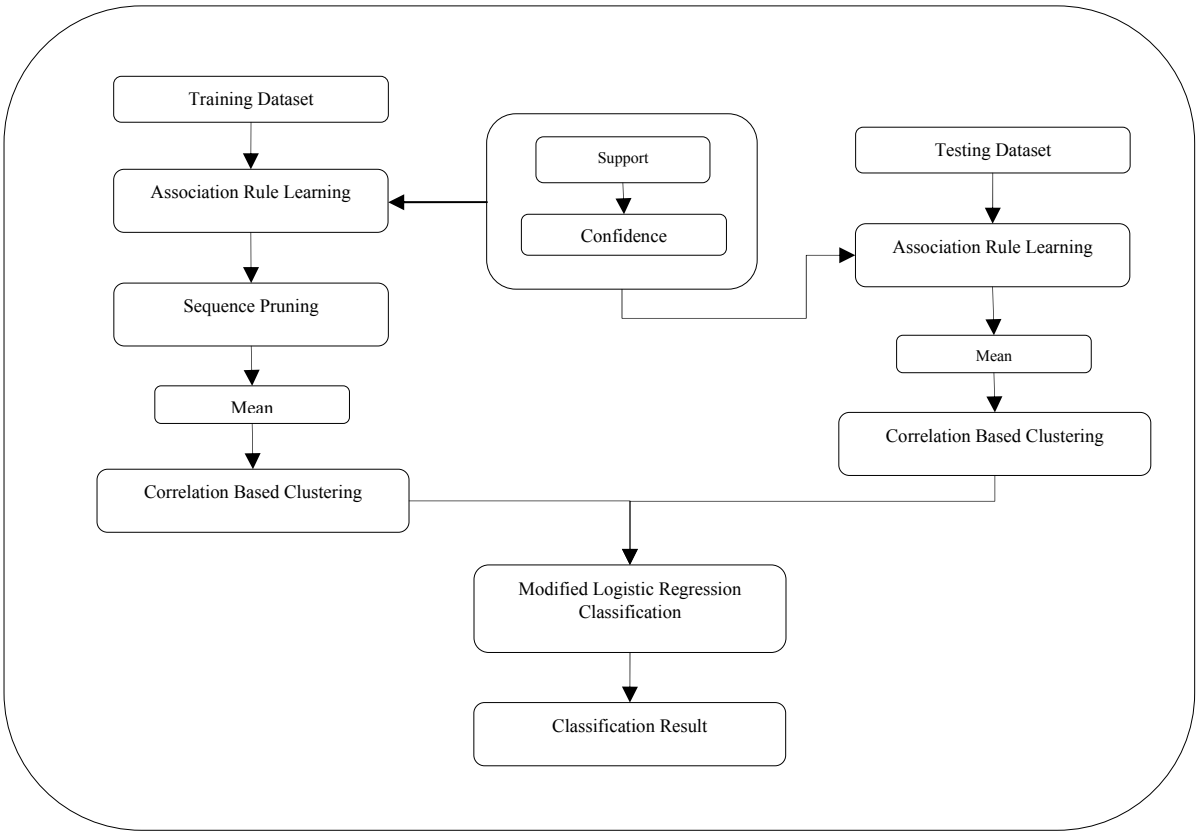

Figure -1: Flow of research

7. Experimental Setup

This section describes about the experimental setup and datasets used for carrying the simulation performance of an incremental Brute-Force approach for clustering DNA Genomic Datasets.

Experimental Setup: To validate the feasibility and performance of the proposed approach, implementation has been done in MATLAB 7.0 (C2D, 2.0 GHz, 2 GB RAM) and applied it to both of real gene expression data and synthetic data.

Datasets: To evaluate the performance of proposed approach, two real and one synthetic DNA gene expression data has been considered for study. These Splice Datasets do not possess referenced (class-labeled) information available to the data.

• Splice Datasets I is a synthetic DNA gene expression data of [10x3] matrix.

• Splice Datasets II represents Splice Dataset and is obtained

The dataset contains 699 instances and each instance is having 9 features. There are 16 data points with missing values. Missing values have been replaced by mean of the feature of column in which missing value was found.

• Datasets III is a real biological microarray data of [118x60] matrix that represents growth inhibition factors when 118 drugs with putatively understood methods of action were applied to the NCI60 cell lines.

Cluster Validation Metrics: Since all the datasets (Datasets I, II and III) considered in this chapter for simulation study do not possess class-labeled information, clustering accuracy cannot be used as a cluster validation metric. In this chapter of the paper, HS Ratio is used as cluster validation metric to validate the quality of clusters. Details of the HS Ratio have been already discussed in this section also. It may be noted that "The quality of cluster V increases with higher homogeneity values within C and inferior division principles between V and other group cluster"

Results and Discussion: Clustering performance has been simulated for different value of classifiers and results obtained were tabulated in Table 1.1. The value of parameter accuracy and ROC is been taken for simulation studies. The code algorithm generates number of cluster automatically. The quality of the clusters formed was assessed using HS Ratio. The simulation results shown in Table 1.2 shows that the proposed classification accuracy approach may perform better compared to various clustering algorithm when the no. of cluster is high.

The table 1.3 shows quality of comparing multi class datasets to get classification accuracy. The datasets formed used was listed in the first column, it is followed by various different algorithms used and the efficiency of the proposed also determined in the given table.

8. Final Results and Discussions

The final results are listed in terms of Clustering performance for splice dataset in table1.1, followed by classification accuracy of proposed algorithm in table 1.2, and comparison of multi class datasets to get classification accuracy in table 1.3.

Table 1.1: Clustering Performance Splice dataset

clustering Performance for Splice dataset		
classifier	**Accuracy**	**ROC**
c4.5	89.25	90.2
naïve Bayes	91.6	92.5
SVM	90.2	91.64
simple Cart	89.54	90.35
K-NN	90.62	91.54
Proposed	92.87	93.12

Table 1.2: Classification accuracy of proposed algorithm

Classification Accuracy							
Top N genes	**UFRFS**	**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

Table 1.3: Comparison of Multi class datasets to get classification accuracy

comparing Multi class Datasets to get classification accuracy						
Dataset	**MOEDA**	**TSP**	**K-TSP**	**GA-ESP**	**KernelPL - KNN**	**Proposed**
Leukemia	1	0.971	0.971	0.965	1	1
SRBCT	0.956	0.95	1	0.98	0.96	0.98
Lung	0.957	0.836	0.94	0.9	0.95	0.97
Splice Dataset	0.96	0.96	0.95	0.95	0.95	0.96

The above tables shows the efficiency of the proposed algorithm, in terms of the performance measures mentioned above

9. Conclusion

The thriving component in various datasets and sequences of DNA data mined with the clustering elements and checking the variety of genome expressions achieved completely. The study procedure is however to accomplish auxiliary objective and advancements in efficiently in its own performance measures, and cross modules with DNA Genomic Database with comparison of multiclass datasets to get accuracy is determined. The preliminary segment of the assignment, called mapping with various datasets, it has disjointed the genetic material into cluster groups as a mutual place of synchronized terminological expressions. Here a huge Data mined processor is been used to point out the position of the group clustered genes and appearance of genes.

References

[1] ENFSI DNA Working Group, DNA-Database Management Review and Recommendations, with financial support from the ISEC Programme, European Commission- Directorate General Justice and Home Affairs April 2012.
[2] Marina Andrade & Manuel Alberto M. Ferreira, Criminal and Civil Identification with DNA Databases Using Bayesian Networks, International Journal of Security, (IJS), Volume (3): Issue (4), PP 65-74, 2010
[3] V.N. Rajavarman and S.P. Rajagopalan, Feature Selection in Data-Mining for Genetics Using Genetic Algorithm, Journal of Computer Science 3 (9):723-725, 2007, ISSN 1549-3636, Science Publications, 2007, PP 723-725
[4] Chan Wai Keung Brian, Data Mining Using Genetic Algorithm, City University of Hong Kong, Dissertation, Hong Kong, August 2006
[5] Yang, J. and V. Honoavar, 2005. Feature Extraction Construction and Selection: A data Mining Perspective, chapter 1: Feature Subset Selection Using a Genetic Algorithm, H. Liu and H. Motoda Eds, massachussetts: kluwer academic publishers Ed., pp: 117-136.
[6] Bates Congdon, C., 2002. A comparison of genetic algorithm and other machine learning systems on a complex classi. Cation task from common disease research. Ph.D Thesis, University of Michigan.
[7] VIJAY ARPUTHARAJ J and Dr.R.MANICKA CHEZIAN , 2013. DATA MINING WITH HUMAN GENETICS TO ENHANCE GENE BASED ALGORITHM AND DNA DATABASE SECURITY .International Journal of Computer Engineering & Technology (IJCET).Volume:4, Issue: 3, Pages: 176-181.
[8] Dr.C.Sunil Kumar,J.Seetha, S.R.Vinotha, Security Implications of Distributed Database Management System Models, International Journal of Soft Computing And Software Engineering (JSCSE),e-ISSN: 2251-7545, Vol.2, No.11, 2012, PP 20-28.
[9] Mount David W., Bioinformatics – Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, 2001.
[10] Rajesh S., Prathima S., Reddy L.S.S., Unusual Pattern Detection in DNA Database Using KMP Algorithm, International Journal of Computer Applications (0975 - 8887)Volume 1 – No. 22, 2010.
[11] Kurzrock R., Kantarjian, H. M. Druker B. J., Talpaz, M. (2003). "Philadelphia chromosome positive leukemias: From basic mechanisms to molecular therapeutics". Annals of internal medicine 138 (10): 819–830.
[12] Pakakasama S., Kajanachumpol S., Kanjanapongkul S., Sirachainan N., Meekaewkunchorn A.,Ningsanond V., Hongeng, S. (2008). "Simple multiplex RT-PCR for identifying common fusion transcripts in childhood acute leukemia". International Journal of Laboratory Hematology 30 (4): 286–291.
[13] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res.2006;34(Database):D16–20.
[14] Manju B R, Dr A R Rajan and Dr V Sugumaran, “Optimizing the Parameters of Wavelets for Pattern Matching using GA”, International Journal of Advanced Research in Engineering & Technology (IJARET), Volume 3, Issue 1, 2012, pp. 77 - 85, ISSN Print: 0976-6480, ISSN Online: 0976-6499.
[15] Vijay Arputharaj J and Dr.R.Manicka Chezian, “A Collective Algorithmic Approach- For Enhanced DNA Database Security”, International Journal of Management and Information technology, Vol4, No1, 2013, ISSN 2278-5612,PP 174-178