International Journal of Engineering & Technology, 7 (4) (2018) 5292-52996

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.25557
Research paper

Correlation-based clustering and the modified naïve-Bayesian-classification for gene-sequence data analysis

Vijay Arputharaj [1] *, Dr.S.Sheeja [2], Dr. K. Anuradha [3]

[1] Research Scholar, Karpagam Academy of Higher Education, Coimbatore
[2] Associate Professor, Department of CS, CA, IT, Karpagam Academy of Higher Education, Coimbatore
[3] Associate Professor Dept. of Computer Applications, Karpagam College of Engineering, Coimbatore
**Corresponding author E-mail: phdvij@gmail.com*

Abstract

Correlation based Clustering separates the statistical data from the most favourable amount of clusters with corresponding to the statistically analysed data points. As we know, Data mining is the technique of figuring out progression of determining patterns inside huge statistics and datasets, which concerns techniques related to connection with machine related learning, statistics and also the advanced database systems. This technique denotes the gene sequence using the novel classification technique, which improves the accuracy of classification under the course of dimensionality. Grouping the gene data using correlation-based clustering will reduce the execution time.

Keywords: *Clustering; Classification; Gene Sequence; Data Analysis.*

1. Introduction

The importance of Data mining technique has large implications to human genetics. A variety of fields in human genetics, where data mining technique is applied includes the following: 1) During discrepancy in individual DNA gene sequence, 2)Variance in collection of codec algorithms, 3)Database information security issues, 4)Revolution susceptibility issues, 5)Genetic parental evaluation issues, 6)Detection differences etc. Especially in our country with massive constraint for database may tend to ease in prevention of defrauds. Such defrauds includes Smart Card (Aadhar) frauds, License card frauds, Passport frauds, Ration Card trickery and other similar fraud. Relatively such divination or visual revelation is available for the same. A few advanced data mining techniques are available to enhance it. These techniques are used to approve, attempt and determine, methods for classifying DNA gene sequence elements. DNA gene sequence base elements are "exons" and "introns", out of which exons are coded DNA gene sequence elements and introns non-coded DNA gene sequence elements. So 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. This is a move ahead in computer knowledge in the meticulous network, which led to "data and sequence explosion".

Currently, the numerous developments in government, health care, education, science and information technology raises the density of information. The large size data in electronic form is 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 a 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 subdivision presents an overview of research carried out on classifications- and their applications to gene sequence data analysis, which are reported in this literature study. This chapter is roughly divided into different sections. The primary part deals with research work carried out on several data analysis and secondary part concentrates on the research methodologies carried out with different clustering techniques used.

DNA gene sequences are elevated-throughput techniques for investigating complex data samples. Mining technique makes it possible to evaluate quickly, competently and precisely in the levels of appearances of DNA genes present in a genetic illustration. The relevance of such techniques in various investigational circumstances generates a number of useful data. However, the major difficulty is while scrutinizing the information. Origins of momentous organic data information from raw gene sequence data are obstructed by the difficulty and immensity of the data. In recent days, numerous numerical methods have been in place to overcome the difficulty associated with gene expression data. A few significant methods have been explained here:

- ENFSI DNA Working Group April 2012: Research 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-testing. This should be done under strict conditions, removing any links to the identity of the DNA owner 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 on 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 recent 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. And even in immigration cases 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 to discover associations between genes-genes and genes-environmental factors. They successfully implemented the classical k-means algorithm without any feature selection. The execution time was very large (over 7500 minutes) and results could not be interpreted (In this the features involved in disease could not be identified). So the feature selection phase was required. With the feature selection, the time of execution of k-means had decreased to 1 minute and the results were exploitable. The clusters obtained with k = 2 and their number of occurrences are discussed here. This showed that the k-means algorithm using results of the Genetic Algorithm was able to construct clusters very closely related to the solution presented in results of the workshop. Moreover this solution had been exactly found 4 times out of 10 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] was that it did not have to know any rules of the problem in advance – the rule will could be found through evolution. This was very useful for a very complex and loosely defined problem. The drawback of genetic algorithm was that the definition of the fitness function could be very complicated sometimes. The fitness function might affect the performance of the process significantly if the complexity of the fitness function increased. It was 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 could not be derived even after countless iteration if the genetic operators were wrongly 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, were curable in less than 50 % of patients. Prognostic models based on pre-treatment characteristics, such as the International Prognostic Index (IPI), were currently used to predict outcome in DLBCL. However, clinical outcome models identified neither the molecular basis of clinical heterogeneity, nor specific therapeutic targets. In this paper analysis had 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 a supervised learning prediction method was applied 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 works and studies

To find the disease prediction, small amounts of genes were used for disease prediction which cost less. If amounts of genes increased, it would also increase the cost. Here, multi objective heuristic algorithm which was also called as 'MOEDA' was proposed. It was an enhanced version of Univariate Marginal Distribution Algorithm. There were two main rules that are applied here. They are 1) Higher (values) and Fewer Rule, 2) Forcibly Decrease (variants) Rule. The first rule was applied for assessing and sorting persons. The second rule was applied to generate potential individuals. After that the 'Support Vector Machine' (SVM) categorization was used for gene classification. The main drawback in this methodology was high computation cost. Cost is increased because of number of iterations.

To overcome the limitations above, it proposed "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 was regarded as the input to the system. The input contained gene sequence, instance name and class labels. The generation of Association rules with the support and confidence measures filtered the gene sequences considerably. Then, the Correlation Based Clustering (CBC) was used to create the clusters. Then, the testing was started by giving testing dataset as input. Association rules were used for testing data with support and confidence calculation. Then correlation-based clustering was applied to the testing data. And finally modified logistic regression classification was used as classification algorithm to find the class labels for the testing gene sequence dataset.

4. Materials and methods

Data mining process and data information retrieval: Data mining was an inspection process of the Information contraption in the DNA Gene sequence Databases. It was also an extensive area of computer pasture with inter discipline way-. It had the calculations and computation improvement to determine prototype in huge information datasets connecting systems at the relationships of reproductions, in understanding mechanism, knowledge related information and database schemes. The general objective of an information data mining process was to generate data in succession from a large dataset. And modernize the generated data to an intense and understandable establishment for supporting idea. For commencing the unrefined assessment rapidity, the above process absorbed 1)catalogs data management uniqueness, 2)Set of data figures and preprocessing, 3)DNA gene data model and Database hypothesis manifestation, 4)Database model metrics, 5)Intricacy alarm, and 6)DNA Database back-processing of revealed foundations, specters etc.

In the major part, data mining and information discovery were the progression of dissect data and information. Initiation of data was devoted to productive statistical DNA gene database information in progression. It also protected to mature, expand process, expenses equally in data mining. Data mining [5] was single methodical tools for considering its information. It also supported data abusers to study data information statistics like innumerable different scales, sub clusters etc. which were used to identify the relations in DNA gene dataset process. Precisely, data mining technique was the development of decision relationship or prototype in several statistics of huge relational DNA gene databases.

Although data information mining method was sensibly an inventive term for organic datasets like DNA Gene databases, the knowledge was not proficient in the particular associated field. With increasing computer techniques like computer dispensation supremacy, computer disk storing liberty, and geometric reckoning the exactness of assessment is developing.

For an instance, one DNA Gene dataset applied in the data mining capacity of database to assess limited succession model. This stimulated to help datasets concepts on exact principles. It also tended to examine and demonstrate the datasets typically by processing data items on it. The following are specific aspects of data mining methods that were used.

Data, Information & Knowledge

Information applied in the DNA gene profiles. such as names, places or any details, statistics, or manuscript that could be developed by a computer. In current scenario, business groups were building up huge and mounting quantity of data in unrelated deals and disparate database datasets. It also comprised an equipped or transactional data like unprocessed values in dataset, non-operational data, such as statistical data, and comprehensive informatics data etc.

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

Information-The samples, transactions, or interaction amid everyone only the specific information can give information about the datasets. For instance, psychoanalysis of DNA principles in point of assessment of parental data comparisons submitted revolved on its developments.

Knowledge- Database information and figures could be rehabilitated into acquaintance data information about precedent previous patterns and prospect tendency. For instance, synopsis information on datasets could be examined in light fundamentals to supply knowledge of scrutiny elements. So, an ultimate user could determine which data were most vulnerable to process constraints.

Data Warehouses- Exaggerated development in DNA dataset along with data information incorporates, indulgence supremacy, information transmission, and store competences. It facilitated to put together, their various databases keen on data information warehouses in process elements. Data warehousing processes were scheduled general information statistic, direction and repossession.

Classes: The accrued information was used to place data in prearranged groups. For example, an alteration development association classes could colliery data to establish when they typically arranged exacting data sets in it. This information might be an access to augment data set swapping by much number of different classes [6].

Clusters: Data materials were bunched or inter-organized units to rational links or customer penchants. [7]

Associations: Data could be exhumed to compose comparative links. This was a major case of associative mining. [8]

Sequential patterns: Data was exhumed to expect recital behavioral samples and penchants.

Individual steps of assessments were available in data information mining.

Artificial and neural networks: These were a nonlinear prophetic facsimile that revision from part to part training and which were like inherited neural complex networks in collections.

Genetic algorithms: several optimization methods that utilized practices such as heritable muddle, transmutation variation, and probable group in an arrangement based on the discernment of ordinary DNA gene-based database advancement. [9][10]

Decision trees: Tree formed as assumptions that indicated number of choices. These sentences originated rule and set of laws for the compilation of a data set.

Data information mining performed to discover knowledge from complete DNA gene Database. It could have few stages in genomic analysis for data mining. The easiest way was an in-depth scrutiny of the outcome from an only inquiry with a genomic explorer. In this stage, an individual might find a gene or indication name, or by planning an order to the genomic explorer. An annoyed assessment of a range of explanation tracks might assist to formulate intellect of the inquiry section. This was a well-accepted utilization of every genome explorer [11]. Data information mining was a reverse of information data retrieval in the usual logic; it was not found on determined criteria; it would expose several veiled patterns by investigating specific data [12].

This proposed advancement was a mixture of confined protection practices overtly digital corroboration and justification practice pursued with an information mining in the DNA Genetic database systems. This research over indenture during the evolution of genetic based algorithm with apposite protection surfaces in DNA genetic gene databases which was Splice Dataset [13 - 15].

5. Performance parameters

We analyze the performance of the proposed work for the 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, followed by sequence pruning, mean, correlation-based clustering which is shown in figure 1.

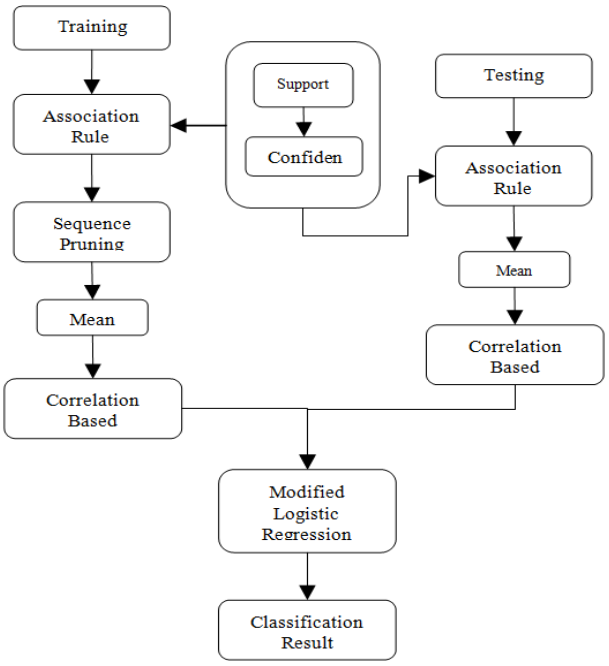

Fig. 1: Flow of Research.

7. Experimental evaluation

This subdivision illustrates the tentative setup (experimental) and data sets applied for conveying the reproduction performance of an incremental Brute Force (BF) approach for group clustering DNA Genetic Datasets.

Experimental Setup: To authenticate the viability and presentation of the planned approach, execution has been done in Java Virtual Machine. In addition to it, actual genetic material expression data and artificial data were used.

Datasets: To assess the recital element of planned advancement, two genuine and one fake DNA genetic material expression data have been measured for learning. These Splice Data sets do not acquire mentioned (class-labeled item) data information existing to the data in dataset.

- Splice Data sets I is an artificial DNA genetic material expression data of matrix.
- Splice Data sets II characterized as real Splice Dataset and is obtained

The dataset is surrounded by around 700 instances and each instance contains around 9-10 features. There are sixteen data positions with absent values. Those mislaid values have been swapped by the facets of discourse in which absent values were established.

- Datasets III is an actual organic array data of matrix that presents the expansion reticence factors while number of drugs with putatively unstated methods of act was practical to the cells.

Cluster Validation Metrics: while every data set -Datasets I, II and III measured in this section for replication research study do not acquire the class-labeled information, clustering exactness cannot be applied as a cluster group substantiation metrics. In this section of the document, H-S Ratio is applied as cluster group corroboration metrics to authorize the quality of cluster groups. Particulars of the H-S Ratio have been previously elaborated in the current part also. It can be distinguished that "The worth of cluster group V enlarges with upper homo genetic values in C and lesser partition values among V and further set cluster groups".

Results and Discussion: 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 much number of clusters repeatedly. The excellence of the cluster groups created was calculated using H-S Ratio statistics. The actual simulation consequences mentioned in Table 1.2 shows that the projected classification accurateness in the proposed approach may present improved contrast to a variety of clustering algorithm when the number of cluster group is very high.

The last 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 discussion

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	UFR FS	Alg 1	UFSF S	UFRD R	FRMI M	CF S	Pro-posed
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 (Consolidated Accuracy with Multi Class)

Comparing Multi class Datasets to get classification accuracy						
Dataset	MOE DA	TS P	K-TSP	GA-ESP	KernelPLS+ KNN	Pro-posed
Leukemia	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 above tables shows the efficiency of the proposed algorithm, in terms of the performance measures mentioned above

9. Conclusion

The clustering essentials were used to establish the sequences of extracted DNA genetic data. Using the data mining technique the diversity of genetic terms was attained absolutely. The existing process is tough to complete supporting intentions and progress efficiently in its own performance measures. Cross modules with DNA Genomic Databases and comparison of multiclass datasets have helped to attain the determined accuracy. The preliminary segment of the assignment, called mapping with various datasets has affected the hereditary objects and split them into clustering groups, as a mutual place of synchronized terminological terms. It is a huge Data mined processor and has been applied to mention the position of the group clustered genetic material and also appearance of genetic material.

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. https://doi.org/10.7326/0003-4819-138-10-200305200-00010.

[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. https://doi.org/10.1111/j.1751-553X.2007.00954.x.

[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.