RESEARCH ARTICLE OPEN ACCESS

Application Of Data Mining In Bioinformatics

Dr A.Chandrabose1 T.Manivannan2 , M.Jayakandan3 T.Sukumar4
Associate.Professor, Assistant professor, Assistant professor, Assistant professor,
Department of computer science
Edayathangudy G.S.Pillay Arts & ScienceCollege Nagapattinam.
Corresponding Author; Dr A..Chandrabose1t

ABSTRACT

With the widespread use of databases and the explosive growth in their sizes, there is a need to effectively utilize these massive volumes of data. This is where data mining comes in handy, as it scours the databases for extracting hidden patterns, finding hidden information, decision making and hypothesis testing. Bioinformatics, an upcoming field in today's world, which involves use of large databases can be effectively searched through data mining techniques to derive useful rules.

keywords: Association Rule Mining: -(a) a priori (b) partitionGenetic algorithms Clustering: - (a) k-means (b) k-medoids lassification Rule Mining: - Decision tree generation using (a) gini index (b) entropy value. clustering

Date Of Submission:05-09-2018 Date Of Acceptance: 21-09-2018

I. INTRODUCTION

1.1 Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

1.2 .The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

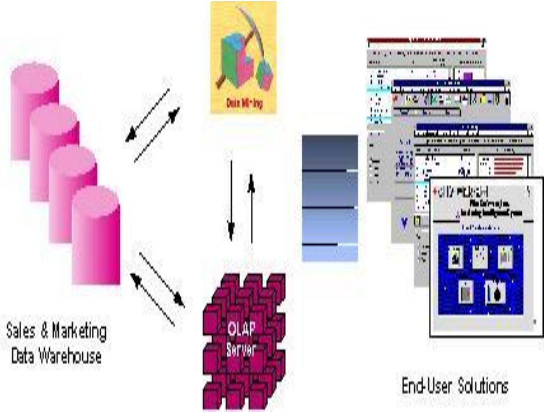


Fig 1.2: An Overview of the Steps Comprising the KDD

The iterative process consists of the following steps:

Data cleaning**:** Also known as data cleansing, it is a phase in which noisy data and irrelevant data are removed from the collection. Data integration**:** At this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection**:** At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data mining**:** It is the crucial step in which clever techniques are applied to extract data patterns potentially useful.

Pattern evaluation**:** In this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation**:** It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

In other applications such as text editors, even simple algorithms for this problem usually suffice, but DNA sequences cause these algorithms to exhibit near-worst case behavior due to their small number of distinct characters. Data sets representing entire genomes' worth of DNA sequences, are difficult to use without annotations, which label the locations of genes and regulatory elements on each chromosome.

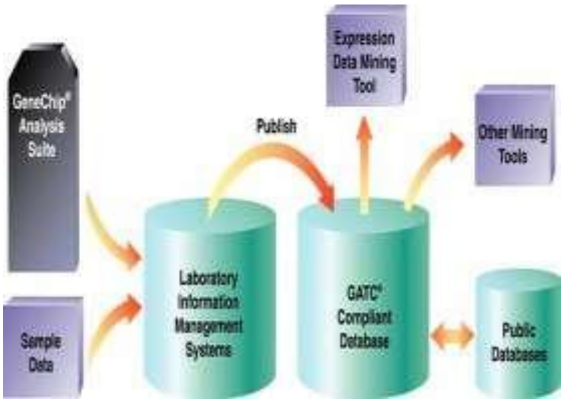


Figure 1.3 DNA sequences

II. ALGORITHM

Let I = {i_1, i_2, … , i_m} be a set of literals, called items. Let D be a set of transactions where

each transaction T is a set of items such that $T \subseteq I$. Let X is a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \Phi$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set

D if s% of transactions in D contains X ∪Y. association rules is decomposed into the following two sub-problems: .

2.1.1 Apriori Algorithm

L_k: - Set of large k-itemsets (those with minimun support).

C_k: - Set of candidate k-itemsets (potentially large itemsets)

The algorithm consists of two phases :- The candidate generation step and pruning step.
1.Candidate generation step (gen_cand_itemsets)
$C_k=\Phi$ for all itemsets $l_1 Є L_{k-1}$ do for all itemsets $l_2 Є L_{k-1}$ do if $l_1[1]=l_2[1]$ ^ $l_1[2]=l_2[2]$ ^…^ $l_1[k-1]=l_2[k-1]$ then c= $l_1[1]$, $l_1[2]$,…, $l_1[k-1]$, $l_2[k-1]$ $C_k= C_k ʋ \{c\}$

2.1.2 Pruning step (Prune)

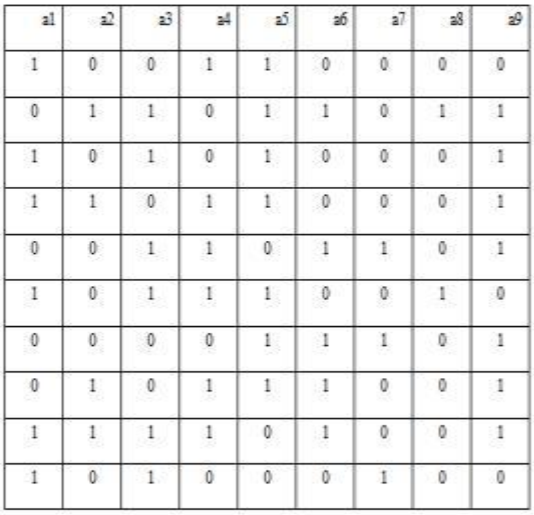
```
For all c $Є C_k$
For all (k-1) subsets d of c do If
        d not belongs to $L_{k-1}$
Then $C_k = C_k \setminus \{c\}$
```

2.1.3Apriori Algorithm

Initialize: k:=1; C_1= all the 1- itemsets; L_1:= {frequent 1-itemsets in D}; while $L_{k-1} \neq \Phi$ do begin C_k := gencand_itemsets with the given L_{k-1} Prune(C_k)for each transaction T in the database do increment the count of all candidate itemsets in C_k that are contained in T; L_k:= candidates in C_k with minimum support; k := k+1;

2.1 Comparison between a priori and partition algorithms

The a priori and partition algorithms were compared using a transaction database. The database represents the sale of 9 items i.e. a1, a2, …, a9 and the rows are the transactions. For eg the first transaction shows that the sale of item 1, item 4 and item 5 have occurred together.



a1	a2	a3	a4	a5	a6	a7	a8	a9
1	0	0	1	1	0	0	0	0
0	1	1	0	1	1	0	1	1
1	0	1	0	1	0	0	0	1
1	1	0	1	1	0	0	0	1
0	0	1	1	0	1	1	0	1
1	0	1	1	1	0	0	1	0
0	0	0	0	1	1	1	0	1
0	1	0	1	1	1	0	0	1
1	1	1	1	0	1	0	0	1
1	0	1	0	0	0	1	0	0

III. CLUSTERING

Clustering is a technique for the purpose of division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. An example of clustering is depicted in Figure 2.1. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful, broadly applicable data mining clustering methods surveyed below.

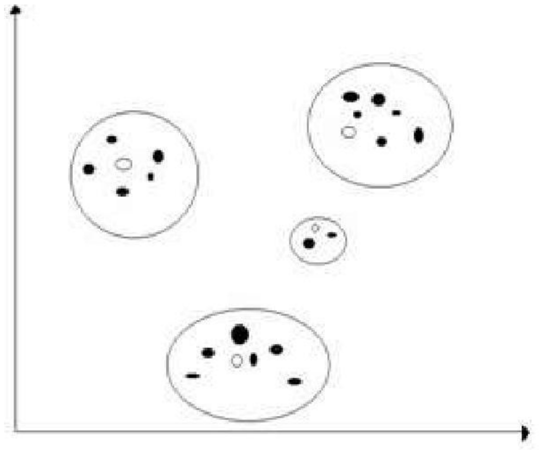


Fig 3.1 Clustering example

IV. ALGORITHM

Input: - Database of objects D

Select arbitrarily k representative objects only from the dataset K_{med}. Mark these objects as selected and the remaining as non-selected do for all selected objects O_i do for all non-selected objects O compute C_{ih} end do end do select imin,hmin such that Cimin,hmin= Mini.h Cih if $C_{imin,hmin}$ <0 then swap; mark O_i as non-selected and O_h as selected repeat find clusters C_1, C_2, $C_{3.....}$,C_k

4.1.K-means Algorithm

It is similar to the k-medoids algorithm except that the centroids are computed as the arithmetic mean of all points of a cluster and the cluster centers are not necessarily objects in the database.

4.2Algorithm

Input: - Database of objects D

Select arbitrarily k representative objects K_{mean} Mark these objects as selected and the remaining as non-selected do for all selected objects O_i do for all non-selected objects O_i compute end do select imin,hmin such that Cimin,hmin= Mini.h Cih if $C_{imin,hmin}$<0 then swap;mark O_i as non-selected and O_h as selected repeat find clusters C_1, C_2, $C_{3.....}$,C_k

V. COMPARISON BETWEEN K-MEANS AND K-MEDOIDS ALGORITHM

The k-means and k-medoids algorithm was compared using a student database.

roll	math_mark	phy_mark	chem_mark	avg
1	70	65	75	70
2	89	69	77	78
3	77	86	88	84
4	86	77	67	77
5	94	92	85	90
6	67	66	70	68
7	55	71	65	64
8	76	88	80	81
9	88	68	77	78
10	60	54	68	61

Table No. 5.1 Student database

The following observations were made:-
Number of interchanges: (i) K-means – 116
(ii) K-medoids – 171

It was found that K-means works better than K-medoids since it performs clustering with less number of interchanges.

We constructed the decision tree using the CART method using

(a) gini index(g_i) as the splitting index where $g_i = 1 - \sum p_i^2$

(b) entropy value(e_i) as the splitting index where $e_i = p_i \log(p_i)$ p_i is the frequency of occurrence of the class i in T

5.2 Comparison between gini index and entropy value as the splitting index

outlook	temp	Humidity	windy	class
sunny	77	75	true	play
rain	65	88	false	no play
sunny	62	80	true	play
overcast	71	78	true	play
rain	59	89	false	no play
overcast	66	82	false	no play
sunny	61	74	true	play
overcast	75	86	false	no play
sunny	59	76	false	no play
rain	80	88	true	play

Table No. 5.2 Classification database

We used the above table to construct a decision tree using gini index and entropy value as the splitting index and got the following results.

Gini index for outlook = 0.49972892 humidity = 0.49980438 windy = 0.49986857 temp = 0.49997646 Time for Execution:-859 ms Entropy value for outlook = 1.9997959 humidity = 1.9999048 windy = 1.9999974 temp = 2.000153 Time for Execution:-875 ms

The best splitting attribute in both the indices was found to be outlook and also the order of splitting attributes were found to be same for both the indices.

5.3 Soft Computing Techniques

Soft Computing refers to a collection of computational techniques in computer science, artificial intelligence, machine learning and some engineering disciplines, which attempt to study, model and analyze very complex phenomena. Earlier computational approaches could model and precisely analyze only relatively simple systems. More complex systems arising in biology, medicine, the humanities, management sciences, and similar fields often remained intractable to conventional mathematical and analytical methods. This is where soft computing provides the solution. Key areas of soft computing include the following::

5.3.1 a. Fuzzy logic

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic. It can be thought of as the application side of fuzzy set theory dealing with well thought out real world expert values for a complex problem. Fuzzy logic can be used to control household appliances such as washing machines (which sense load size and detergent concentration and adjust their wash cycles accordingly) and refrigerators.

5.3.2 b.Neura Networks

Artificial neural networks are computer models built to emulate the human pattern recognition function through a similar parallel processing structure of multiple inputs. A neural network consists of a set of fundamental processing elements (also called neurons) that are distributed in a few hierarchical layers. Most neural networks contain three types of layers: input, hidden, and output. After each neuron in a hidden layer receives the inputs from all of the neurons in a layer ahead of it (typically an input layer), the values are added through applied weights and converted to an output value by an activation function (e.g., the Sigmoid function). Then, the output is passed to all of the neurons in the next layer, providing a feed- forward path to the output layer.

The algorithm operates through a simple cycle as shown in fig 5.3.3

1. Creation of a population of strings
2. Evaluation of each string
3. Selection of the best strings
4. Genetic manipulation to create a new population of strings.

The GA Maps strings of numbers to each potential solution. Each solution becomes an individual in the population and each string becomes a representation of an individual.

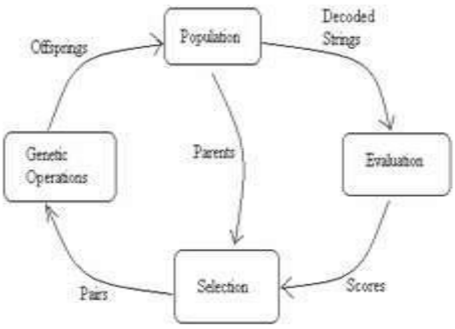


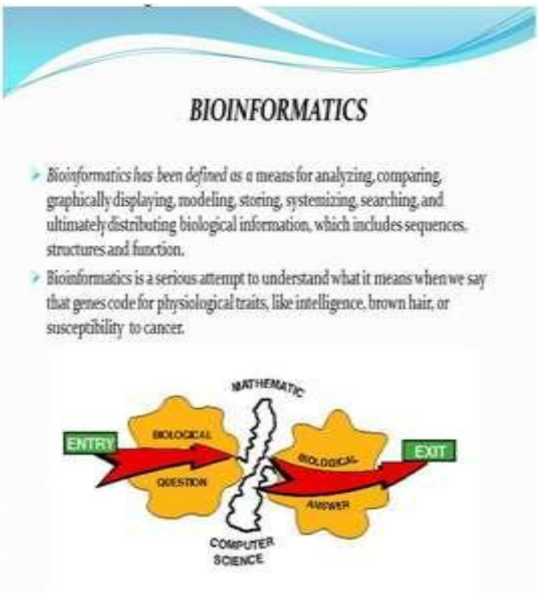
Fig: 5.3.3 : The Genetic Algorithm cycle

Genetic algorithm was successfully applied to classification and association rule mining and the decision tree rules and maximal frequent set found out were more appropriate. The results were obtained in lesser time .

VI. BIOINFORMATICS

6.1 Bioinformatics

6.1.2 .Bioinformatics ^{i}/ˌbaɪ.oʊˌɪnfərˈmætɪks/ is an interdisciplinary field that develops methods and software **tools** for understanding biological data. As an interdisciplinary field of science, **bioinformatics** combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. **bioinformatics** is the application of computer technology to get the information that's stored in certain types of biological data. . Main goal is to convert multitude of complex data into useful information and knowledge.



BIOINFORMATICS

- Bioinformatics has been defined as a means for analyzing, comparing, graphically displaying, modeling, storing, systemizing, searching, and ultimately distributing biological information, which includes sequences, structures and function.
- Bioinformatics is a serious attempt to understand what it means when we say that genes code for physiological traits, like intelligence, brown hair, or susceptibility to cancer.

figure 6.2 The Bioinformatics with Entry and Exit

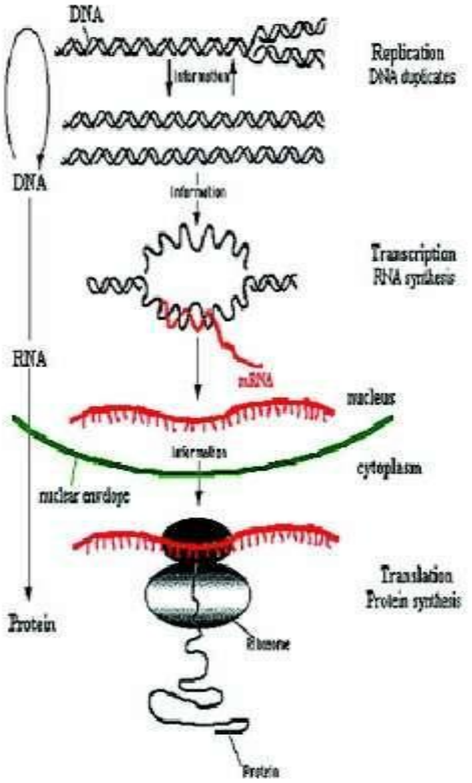


Figure 6.3 DNA ,RNA and Protein

The simplified version of the central dogma is as shown in the diagram. It consists of the following steps: Replication of DNA, Transcription of DNA to RNA, Translation to proteins and protein folding.

7.Replication In the process of DNA replication, all the information in the double-stranded sequence of a DNA helix is duplicated on each strand. This reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms.

7.1 Transcription In this process, the DNA is transcribed to single-stranded nuclear RNA (nRNA) which is then processed to form mature messenger RNA (mRNA).Small nuclear RNA(snRNA) is involved in the maturation process, which includes excising the introns (non-coding).

8. Microarray Experimental Analysis

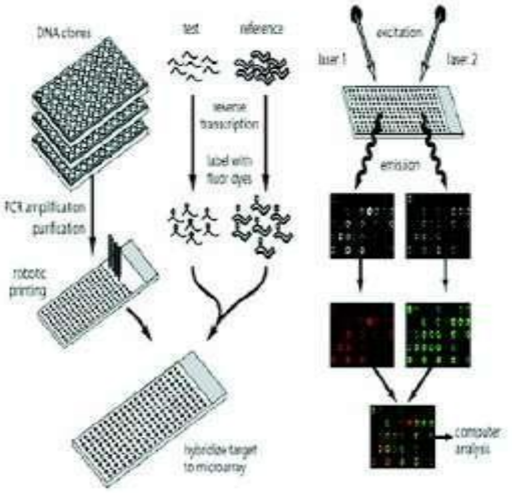
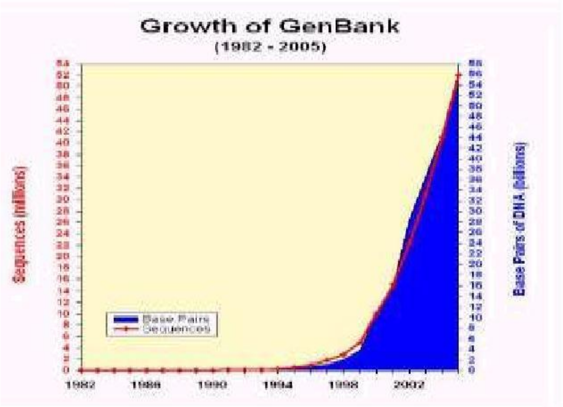


Fig 8.: Microarray experimental analysis

DNA microarrays are created by robotic machines that arrange minuscule amounts of hundreds or thousands of gene sequences on a single microscope slide. Researchers have a database of over 40,000 gene sequences that they can use for this purpose. When a gene is activated, cellular machinery begins to copy certain segments of that gene. The mRNA produced by the cells bind to the original portion of the DNA strand from which it was copied. To determine which genes are turned on and which are turned off in a given

VII. .IMPLEMENTATION



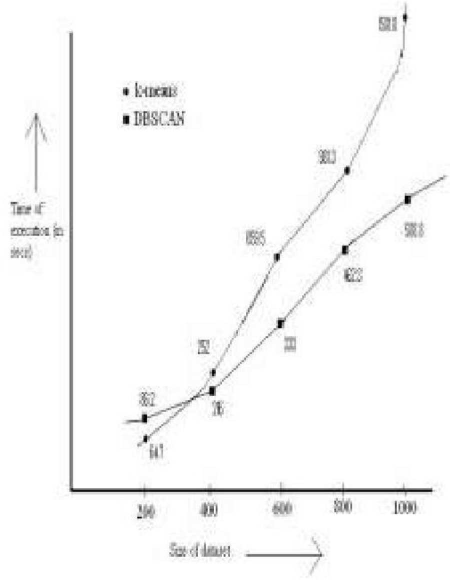


Fig 9.1: Performance comparison of K-means and DBSCAN based on time of execution

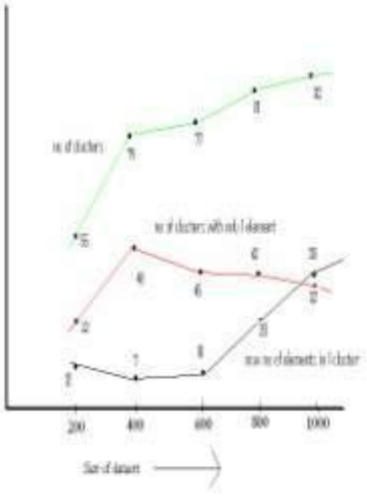


Fig 9.2: Performance of DBSCAN with increasing database size.

For the microarray dataset, it was found that DBSCAN is less efficient than k-means when the database is small but for larger database DBSCAN is more accurate and efficient in terms of no. of clusters and time of execution. DBSCAN execution time increases linearly with the increase in database and was much lesser than that of k-means for larger database

.

VIII. CONCLUSION AND FUTURE WORKS

In this report an in-depth study of the varied data mining techniques was made. It was shown how genetic algorithms can be used to optimize the data mining algorithms. The report then gives an introduction to molecular biology and bioinformatics. Then the microarray experimental analysis was studied and the clustering techniques were applied to mine the microarray data. The report basically underlines the role of clustering analysis to cluster genes into groups of similar character. The microarray experiment produces thousands of samples for each gene, clustering can be effectively used to group these genes into disease causing genes and normal genes and to study the various characteristics of different genes under different conditions. This can be used for drug treatment considering the response of the genes to drugs paving the way for diagnosis of incurable diseases like AIDS, Alzheimer's disease.

It can also be used to identify the mechanisms underlying biological processes such as growth and ageing and to track the process of our evolution. Data mining is therefore an effective technique to solve the problem of enormous data faced by researchers in their quest to solve the puzzles of our life.

REFERENCES

[1]. Pujari, Arun . Data Mining Techniques. Nancy: Universities Press, 2001.

[2]. Zhang , Dongsang and Zhou, Lina. ―Data Mining Techniques in Financial Application‖.

[3]. IEEE Transactions on Systems, Man and Cybernetics – Part C : Applications and Reviews, Vol – 34, No- 4, Nov-2004, pp. 513 – 522.

[4]. Han, Jiawei and Kamber, Micheline. Data mining: concepts and techniques, San

[5]. Francisco: Morgan Kaufmann Publishers Inc., CA, 2000

[6]. Piateski, Gregory and Frawley, William. Knowledge Discovery in Databases, Cambridge:

[7]. MIT Press, MA, 1991

[8]. Edder, John F and Abbott , Dean. ―A Comparison of Leading Data Mining Tools.‖

Fourth International Conference on Knowledge discovery and Data Mining, Friday, Aug 28, 1998, NY, pp 19-25.

[9]. Agrawal, R., Imielinski, T. and Swami, A. ―Mining associations between sets of items in massive databases.‖ ACM SIGMOD International Conference on Management of Data, pp 207--216, Washington, DC, May 1993.

[10]. Hipp, Jochen , Guntzer, Ullrich and Nakhaeizadeh, Gholamreza, ―Algorithms for Association Rule Mining – A general Survey and Comparison‖. SIGKDD explorations, Vol 2, Issue – 1, pp 58 – 63, Mar – 2004.

[11]. Agrawal, Rakesh and Srikant, Ramakrishnan. ―Fast Algorithms for Mining Association

[12]. Rules in Large Databases‖ , Proceedings of the 20th International Conference on Very Large

[13]. Data Bases, pp.487-499, September 12-15, 1994

[14]. Gyorodi, Robert S . ―A Comparative Study of Iterative Algorithm in Association Rules Mining‖, Studies in Informatics and Control, Vol – 12, No – 3, pp 205 – 212, Sept 2003.

[15]. Jain , A. K. , Murty, M. N. and Flyn, P. J., ―Data Clustering : A Review‖, ACM Computing Surveys, Vol – 31, No – 3, pp 264 – 323, 1999.

[16]. Salem, S. A. and Nandi, A. K., ―New Assessment Criteria for Clustering Algorithms‖, IEEE Workshop on Machine Learning for Signal Processing, Vol – 7803 – 9518, Sept 2005, pp 285 – 290.

[17]. Cohen, W., ―Fast Effective Rule Induction‖, In Proceedings of the 12th International Conference on Machine Learning, pp 115 – 123, 1995.

[18]. Kamber, M., Winstone, L., et al. ―Generalization and decision tree induction: efficient classification in data mining‖. Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97) High Performance Database Management for Large-Scale Applications, pp.111, April 07-08, 1997

[19]. Goldberg, David E. Genetic Algorithms in Search, Optimization and Machine Learning.

[20]. Boston: Addison-Wesley Longman Publishing Co., 1989

[21]. Punch W. F., Pei M., et al ―Further Research on Future Selection and Classification using Genetic Algorithms‖, 5th International Conference on Genetic Algorithm, Champaign IL, pp 557 – 564, 1993.

[22]. Carvalho, Deborah R. and Frietas, Alex A., ―A Hybrid Decision Tree / Genetic Algorithm Method for Data Mining ‖, Evolutionary Computation I(2), pp 101 – 125, MIT Press, 1993.

[23]. Tsuchiya, T., Ishihara, S., Matsubara, Y., Nishiro, T., Magamachi, M, ―A Method of Learning Decision Tree Using Genetic Algorithm and its Application to Kansei Engineering System‖, IEEE SMC ' 99 Conference Proceedings, Vol – 6, pp 279 – 283, Oct – 99.

[24].Bergeron, Bryan. Bioinformatics Computing. New Delhi: Pearson Education,2003..

[25]. Luscombe, N.M., Greenbaum, D. and Gerstein, M : ―What is Bioinformatics? A Proposed Definition and Overview of the Field.‖.Methods of Information in Medicine,40(4),pp 346-358,May 2001

[26]. Piatetsky-Shapiro, G. and Tamayo, P : ―Microarray Data Mining: Facing the Challenges.‖ SIGKDD Explorations, 5(2), pp 1-5, 2003 .

[27]. Liu, L., Yang, Jiong. and Tung, Anthony. ―Data Mining Techniques for

[28]. Microarray Datasets.‖ Proceedings of the 21st International Conference on Data 13

[29]. Engineering 2005 IEEE. pp 182-192 ,2005

[30]. Shah, Shital C. and Kusiak, Andrew, ― Data Mining and Genetic Algorithm Based Gene Selection‖, Artificial Intelligence in Medicine 2004, (31), pp 183-196, Vol – 2139, 2004.

Dra.Chandrabose1T "Application Of Data Mining In Bioinformatics " International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.9, 2018, pp 46-52