Role of Clustering on Gene Data

Mayilvaganan.M[#1], Hemalatha R[*2]
[#]*Associate Professor, Department of Computer Science, Bharathiyar University*
PSG college of arts and science, Coimbatore, Tamil Nadu, India
[*]*Assistant Professor, Department of computer science, Tiruppur kumaran college for women*
Tiruppur, Tamil Nadu, India.

***Abstract*—** This Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Clustering algorithm used to find groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups .Here this algorithm is applied to find no of occurrences for the gene dataset. After that T is replaced by U. Comparisons are made based on the Execution time and memory efficiency in finding frequent patterns. The performance is analysed based on the different no of instances and confidence in gene data set. The occurrences for modified data and original data are compared together to find cluster structure.

***Keywords*—**Cluster algorithm.

I. INTRODUCTION

Clustering can be considered the most important *unsupervised learning* technique; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. Clustering is "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters[4]-[10]. The dendrogram is a visual representation of the spot correlation data. The individual spots are arranged along the bottom of the dendrogram and referred to as leaf nodes. Spot clusters are formed by joining individual spots or existing spot clusters with the join point referred to as a node. This can be seen in the diagram above. At each dendrogram node we have a right and left sub-branch of clustered spots. In the following discussion, spot clusters can refer to a single spot of a group of spots. The vertical axis is labelled distance and refers to a distance measure between spots or spot clusters. The height of the node can be thought of as the distance value between the right and left sub-branch clusters.

A. Clustering Algorithm

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modelling puts clustering in ahistorical 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[11]. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis. [2].

B. Data for Research

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy. This data set contains 22 attributes and 8124 instances.

C. Original Data

@data
+,S10,t,a,c,t,a,g,c,a,a,t,a,c,g,c,t,t,g,c,g,t,t,c,g,g,t,g,g,t,t,
a,a,g,t,a,t,g,t,a,t,a,a,t,g,c,g,c,g,g,g,c,t,t,g,t,c,g,t
+,AMPC,t,g,c,t,a,t,c,c,t,g,a,c,a,g,t,t,g,t,c,a,c,g,c,t,g,a,t,t
,g,g,t,g,t,c,g,t,t,a,c,a,a,t,c,t,a,a,c,g,c,a,t,c,g,c,c,a,a
+,AROH,g,t,a,c,t,a,g,a,g,a,a,c,t,a,g,t,g,c,a,t,t,a,g,c,t,t,a,t
,t,t,t,t,t,g,t,t,a,t,c,a,t,g,c,t,a,a,c,c,a,c,c,c,g,g,g,c,g
+,DEOP2,a,a,t,t,g,t,g,a,t,g,t,g,t,a,t,c,g,a,a,g,t,g,t,g,t,t,g,
c,g,g,a,g,t,a,g,a,t,g,t,t,a,g,a,a,t,a,c,t,a,a,c,a,a,a,c,t,c
+,LEU1_TRNA,t,c,g,a,t,a,a,t,t,a,a,c,t,a,t,t,g,a,c,g,a,a,a,
a,g,c,t,g,a,a,a,a,c,c,a,c,t,a,g,a,a,t,g,c,g,c,c,t,c,c,g,t,g,g,t,
a,g
+,MALEFG,a,g,g,g,g,g,c,a,a,g,g,a,g,g,a,g,g,a,t,g,g,a,a,a,g,a,g,
g,t,t,g,c,c,g,t,a,t,a,a,a,g,a,a,a,c,t,a,g,g,a,g,t,c,c,g,t,t,t,a,g,g,
t
+,MALK,c,a,g,g,g,g,g,g,t,g,g,a,g,g,a,t,t,t,a,a,g,c,c,a,t,c,t,
c,c,t,g,a,t,g,a,c,g,c,a,t,a,g,t,c,a,g,c,c,c,a,t,c,a,t,g,a,a,t
+,RECA,t,t,t,c,t,a,c,a,a,a,a,c,a,c,t,t,g,a,t,a,c,t,g,t,a,t,g,a,
g,c,a,t,a,c,a,g,t,a,t,a,a,t,t,g,c,t,t,c,a,a,c,a,g,a,a,c,a
+,RPOB,c,g,a,c,t,t,a,a,t,a,t,a,c,t,g,c,g,a,c,a,g,g,a,c,g,t,c,
c,g,t,t,c,t,g,t,g,t,a,a,a,t,c,g,c,a,a,t,g,a,a,a,t,g,g,t,t,t
+,RRNAB_P1,t,t,t,t,a,a,a,t,t,t,c,c,t,c,t,t,g,t,c,a,g,g,c,c,g,
g,a,a,t,a,a,c,t,c,c,c,t,a,t,a,a,t,g,c,g,c,c,a,c,c,a,c,t,g,a,c,a+
,RRNAB_P2,g,c,a,a,a,a,a,t,a,a,a,t,g,c,t,t,g,a,c,t,c,t,g,t,a,
g,c,g,g,g,g,a,a,g,g,c,g,t,a,t,t,a,t,g,c,a,c,a,c,c,c,c,g,c,g,c,c,g
+,RRNDEX_P2,c,c,t,g,a,a,a,t,t,c,a,g,g,g,t,t,g,a,c,t,c,t,g,
a,a,a,g,a,g,g,a,a,a,g,c,g,t,a,a,t,a,t,a,c,g,c,c,a,c,c,t,c,g,c,g
,a,c
+,RRND_P1,g,a,t,c,a,a,a,a,a,a,a,t,a,c,t,t,g,t,g,c,a,a,a,a,a,a
,a,t,t,g,g,g,a,t,c,c,c,t,a,t,a,a,t,g,c,g,c,c,t,c,c,g,t,t,g,a,g,a
+,RRNE_P1,c,t,g,c,a,a,t,t,t,t,t,c,t,a,t,t,g,c,g,g,c,c,t,g,c,g
,g,a,g,a,a,c,t,c,c,c,t,a,t,a,a,t,g,c,g,c,c,t,c,c,a,t,c,g,a,c,a

+,RRNG_P1,t,t,t,a,t,a,t,t,t,t,t,c,g,c,t,t,g,t,c,a,g,g,c,c,g,g,a,a,t,a,a,c,t,c,c,c,t,a,t,a,a,t,g,c,g,c,c,a,c,c,a,c,t,g,a,c,a
+,RRNG_P2,a,a,g,c,a,a,a,g,a,a,a,t,g,c,t,t,g,a,c,t,c,t,g,t,a,g,c,g,g,g,a,a,g,g,c,g,t,a,t,t,a,t,g,c,a,c,a,c,c,g,c,c,g,c,g,c,c
+,RRNX_P1,a,t,g,c,a,t,t,t,t,t,c,c,g,c,t,t,g,t,c,t,t,c,c,t,g,a,g,c,c,g,a,c,t,c,c,c,t,a,t,a,a,t,g,c,g,c,c,t,c,c,a,t,c,g,a,c,a
+,TNAA,a,a,a,a,c,a,a,t,t,t,c,a,g,a,a,t,a,g,a,c,a,a,a,a,a,a,c,t,c,t,g,a,g,t,g,t,a,a,t,a,a,t,g,t,a,g,c,c,t,c,g,t,g,t,c,t,t,g,c
+,TYRT,t,c,t,c,a,a,c,g,t,a,a,c,a,c,t,t,t,a,c,a,g,c,g,g,c,g,c,g,t,c,a,t,t,t,g,a,t,a,t,g,a,t,g,c,g,c,c,c,c,g,c,t,t,c,c,c,g
+,ARAC,g,c,a,a,a,t,a,a,t,c,a,a,t,g,t,g,g,a,c,t,t,t,t,c,t,g,c,c,g,t,g,a,t,t,a,t,a,g,a,c,a,c,t,t,t,t,g,t,t,a,c,g,c,g,t,t,t
+,LACI,g,a,c,a,c,c,a,t,c,g,a,a,t,g,g,c,g,c,a,a,a,a,c,c,t,t,t,c,g,c,g,g,t,a,t,g,g,c,a,t,g,a,t,a,g,c,g,c,c,c,g,g,a,a,g,a,g
+,MALT,a,a,a,a,a,c,g,t,c,a,t,c,g,c,t,t,g,c,a,t,t,a,g,a,a,a,g,g,t,t,t,c,t,g,g,c,c,g,a,c,c,t,t,a,t,a,a,c,c,a,t,t,a,a,t,t,a
+,TRP,t,c,t,g,a,a,a,t,g,a,g,c,t,g,t,t,g,a,c,a,a,t,t,a,a,t,c,a,t,c,g,a,a,c,t,a,g,t,t,a,a,c,t,a,g,t,a,c,g,c,a,a,g,t,t,c,a
+,TRPP2,a,c,c,g,g,a,a,g,a,a,a,a,c,c,g,t,g,a,c,a,t,t,t,t,a,a,c,a,c,g,t,t,t,g,t,t,a,c,a,a,g,g,t,a,a,a,g,g,c,g,a,c,g,c,c,g,c
+,THR,a,a,a,t,t,a,a,a,a,t,t,t,t,a,t,t,g,a,c,t,t,a,g,g,t,c,a,c,t,a,a,a,t,a,c,t,t,t,a,a,c,c,a,a,t,a,t,a,g,g,c,a,t,a,g,c,g
+,BIOB,t,t,g,t,c,a,t,a,a,t,c,g,a,c,t,t,g,t,a,a,a,c,c,a,a,a,t,t,g,a,a,a,a,g,a,t,t,t,a,g,g,t,t,t,a,c,a,a,g,t,c,t,a,c,a,c,c

E. Data Occurrences

NAME	SEQUENCES		NO.OF OCCURENCES		% OF OCCURRENCES	
Alanine	CCG	UCG				
Arginine	CGC	AGA	732		1.0	
Arparagine	ACA	AGA	892		1.3	
Aspartic	GUA					
Cysteine	UUG					
Glumotic Acid	GAA					
Glutamine	ACA	AGA	892		1.3	
Glycine	UGG	AGG				
Histidine	ACC					
Isoleucine	UAU	UAA	389		0.5	
Leucine	UUG	UUC				
Lysine	AAA	AAG	21456		31.5	
Methionine	UGA					
Phenylalanine	UUU		41253		60.5	
Proline	CCC	CCU	411		1.1	
Serine	UCU	UCC				
Threonine	CAC	CAA	367	641	0.5	.94
Tryptophan	CGU					
Tyrosine	AUU					
Valine	UGU		568		0.8	
Adenine	A		8345		12.2	
Guanine	G		9156		13.4	
Cytosine	C		7312		10.7	
Thymine	T		4262		6.2	

II. METHODOLOGY

The proposed methodology is using gene dataset for mining. The proposed data and outputs are taken to find the occurrences. The occurrences are applied in dendogram to get a structure for the output.. Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters[12]-[15]. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type.

III. IMPLEMENTATION

Implementation is a stage, which is crucial in the life cycle of the new system designed. It is the process of changing from the old system to new one. In the existing research work association rule mining is performed in Gene databases. But in proposed clustering algorithm is used based on dendogram method.. Pre-processing is nothing but data cleaning. The unnecessary information is removed or reconfigures the data to ensure a consistent format. Data can be modified or changed into different formats. Cluster analysis is thus a tool of discovery. It may reveal associations and structure in data which, though not previously evident, nevertheless are sensible and useful once found. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as a taxonomy for related animals, insects or plants; or suggest statistical models with which to describe populations; or indicate rules for assigning new cases to classes for identification and diagnostic purposes; or provide measures of definition, size and change in what previously were only broad concepts; or find exemplars to represent classes.

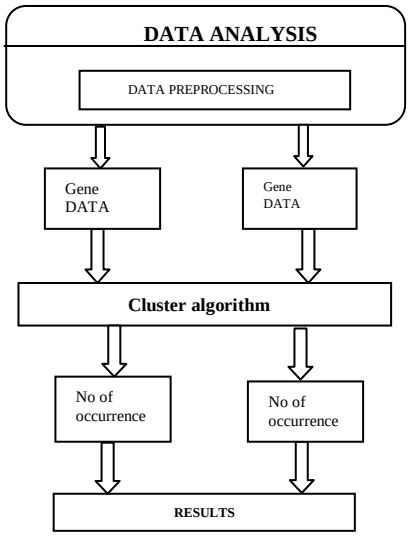

Figure1 Process Flow Diagram

Figure 1 shows the process flow of the proposed system. The performance of proposed work is measured with the existing techniques. Apriori algorithm is applied using association rule mining technique. The Count and position of gene sequences are retrieved using Apriori algorithm. This algorithm is applied separately in string and numerical data. Memory efficiency is calculated and comparisons are made based on it.

A. Conversion Method

X = 24	C = 3	F = 6	U = 21
S = 19	K = 11	N = 14	E = 5
T = 20	W = 23	P = 16	O = 15

Figure2 Conversion table for mushroom data

Figure 2 shows the conversion table for mushroom data. These values are taken from the conversion table which is showed below. Using this value the string data of mushroom data are converted into numerical data. Figure2 shows the string values which is taken from the mushroom data set for example.

IV. RESULTS AND DISCUSSION

The figure 4 shows the conversion table which is used in this paper for converting string data to numerical data .The following Figure3 shows the memory occupied by the string data and numerical data based on apriori algorithm which we discussed above. In this graph, y axis represents the no of items and x axis represents the memory level. The blue line indicates the memory efficiency level of numerical data and the green line indicates the memory efficiency level of string data. Finally the comparisons are made and based on the comparison we can conclude that the memory level of Numerical data is very less based on memory for various no of instances.

Conversion Table

A	= 1	K	= 11	U	= 21
B	= 2	L	= 12	V	= 22
C	= 3	M	= 13	W	= 23
D	= 4	N	= 14	X	= 24
E	= 5	O	= 15	Y	= 25
F	= 6	P	= 16	Z	= 26
G	= 7	Q	= 17		
H	= 8	R	= 18		
I	= 9	S	= 19		
J	= 10	T	= 20		

Figure 3 Conversion table

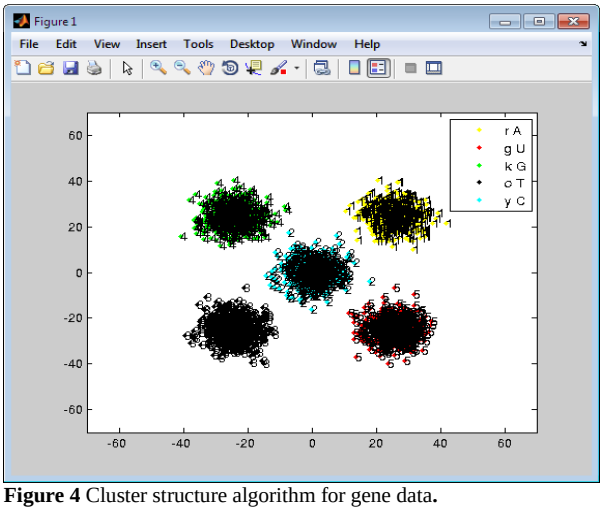

Figure 4 Cluster structure algorithm for gene data**.**

V. CONCLUSION

In this paper we find that the clustering algorithm can be applied both on string and numerical data. Both the values are taken from gene data set ,to calculate the memory efficiency. Though both the output no of occurrences are formulated. Finally dendogram used to find out the structure any other algorithms to find out the time and memory efficiency.

REFERENCES

[1] Role of Association Rule Mining in Numerical and numerical Data. Bhavani K, Hemalatha.R

[2] Survey of clustering data mining techniques. Pavel Berkhin.

[3] Comparison between clustering algorithm. Osama abu abbas.

[4] Supervised clustering algorithm and benefits. ChristopF.Eick.

[5] Bayardo, Roberto J., Jr.; Agrawal, Rakesh; Gunopulos, Dimitrios (2000). "Constraint-based rule mining in large, dense databases". *Data Mining and Knowledge Discovery* (2): 217–240. Doi:10.1023/A:1009895914772.

[6] Webb, Geoffrey I. (2000); Efficient Search for Association Rules, in Ramakrishnan, Raghu; and Stolfo, Sal; eds.; Proceedings of the Sixth ACM SIGKDD International *Conference on Knowledge Discovery and Data Mining (KDD-2000), Boston, MA*, New York.

[7] http://www.b3intelligence.com/NumericalDataMinig.html

[8] http://en.wikipedia.org/wiki/Numerical_analysis

[9] http://www.saedsayad.com/zeror.html

[10] http://www.cogsys.wiai.unibamberg.de/teaching/ss05/ml/slides/cogsysII-6.pdf

[11] http://www.slideshare.net/totoyou/covering-rulesbased-algorithm

[12] M.Anandavalli , M.K.Ghose , K.Gouthaman ,"Association Rule Mining in Genomics",International journal of computer Theory and engineering ,Vol.2,No.2 April,2010.

[13] Arun.K.Pujari"data mining techniques ",Universities Press (india) private limited.2001.

[14] F.Braz,"A review of the association rules data mining techniques for the analysis of gene expressions"

[15] Douglas Trewartha, "Investigating data mining in MATLAB ",Rhodes University 2006.