Role of Association Rule Mining in String and Numerical Data

Bhavani K[#1], Hemalatha R[*2]

[#]*Research Scholar, 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. Association rule mining discovers correlations between different item sets in a transaction database. The proposed methodology is implemented in Apriori algorithm. The given string data is applied with apriori algorithm and the memory efficiency is calculated to get the output measures. The same string data are assigned with numerical values and the apriori algorithm is applied on the numerical data. Based on the numerical values to get the measures. 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 mushroom data set.

***Keywords*—**Association Rule mining, Apriori algorithm, Numerical data analysis .

I. INTRODUCTION

Association rule mining is one of the classical data mining processes, which finds associated item sets from a large number of transactions. Association rule mining is comes under Data mining techniques. In this paper we use this technique to search and mine the very large database. Association rule mining is the discovery of association relationships or correlations among a set of items. Each association rule is in the form of "X => Y" where X and Y are two item sets, which means if X is in a transaction, Y is probably in the same transaction as well. The process of finding association rules can be divided into two steps. First, the set of frequent item sets are computed. Then, the set of association rules can be generated from the set of frequent item sets. While the latter problem is computationally , inexpensive the problem of mining frequent item sets has an exponential time complexity and it is very costly.

A. Apriori Algorithm

Association rule mining is one of the classical data mining processes, which finds associated item sets from a large number of transactions. Apriori discovers patterns with frequency above the minimum support threshold. Therefore, in order to find associations involving rare events, the algorithm must run with very low minimum support values. The Apriori algorithm calculates rules that express probabilistic relationships between items in frequent item sets [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,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,c,a,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,a,g,t,c,c,g,t,t,t,a,g,g,
t
+,MALK,c,a,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,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,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

II. METHODOLOGY

The proposed methodology is using mushroom dataset for mining. Association rule mining is one of the classical data mining processes, which finds associated item sets from a large number of transactions. FIM-Frequent Item Mapping is an important data mining problem which detect frequent item sets in a gene database. By mining frequent patterns we can easily identify the defects occurred; and can rectify it. This can be done with the help of Apriori algorithm. The proposed system can be solved to achieve the effect of existing algorithms for mining. The string mushroom data set is converted into numerical data and apriori algorithm is applied in both the data sets. Finally the memory efficiency is calculated and comparison is made based on time and memory.

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 proposed research work association rule mining is performed in Gene databases. The most efficient Apriori algorithm is implemented using Matlab tool. 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. The Apriori algorithm uses indexed data for generating sequence sets and frequent item sets are identified from gene database.

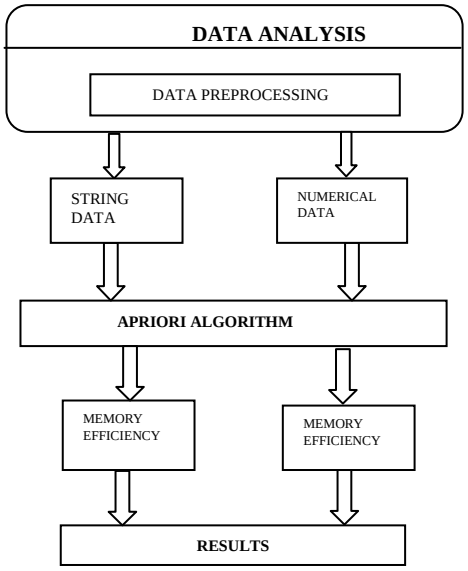

Figure1 Process Flow Diagram

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

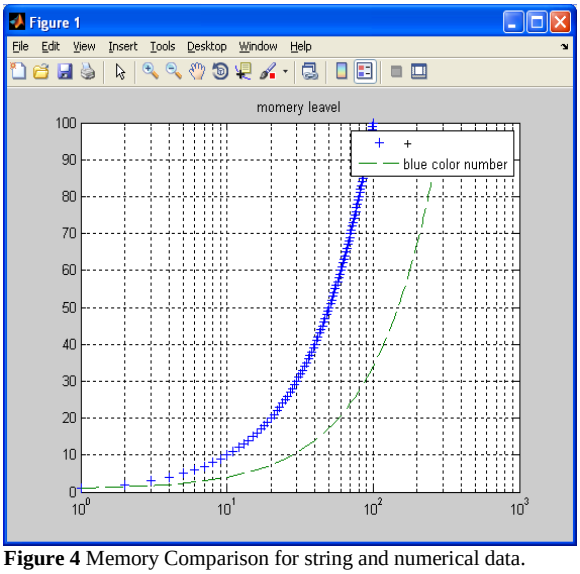

Figure 4 Memory Comparison for string and numerical data.

V. CONCLUSION

In this paper we find that the apriori algorithm can be applied both on string and numerical data. Both the values are taken from mushroom data set ,to calculate the memory efficiency. Though both the output are formulated, it is the numerical data that has best memory efficiency than string data. In future, the same method can be applied using any other algorithms to find out the time and memory efficiency.

REFERENCES

[1] Role of Association Rule Mining in Numerical Data Analysis Sudhir Jagtap, Kodge B. G., Shinde G. N., Devshette P. M

[2] M.Anandavalli, M.K.Ghose ,K.Gauthaman,"Association Rule Mining in Geonomics",International journal of Computer Theory and Engineering Vol.2 ,No.2 April 2010.

[3] Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Cambridge, MA.

[4] Role of association rule mining in numerical data analysis, sudhir Sudhir Jagtap, Kodge B. G., Shinde G. N., Devshette P. M

[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.