© 2014 IJIRT | Volume 1 Issue 11 | ISSN: 2349-6002

Literature Review about Emerging Pattern and Frequent Pattern Growth algorithm apply on gene Expression data

Dinkal Shah[1], Narendrasinh Limbad[2]
[1]*ME Student, Dept. of IT*
[2]*Asst. Professor, Dept. of IT*
LJIET, Ahmedabad, Gujarat, India

***Abstract*- Data mining is a extract knowledge from data. KDD find the interesting or useful pattern and relationship of patterns. There are two types of pattern described (1) Emerging Pattern (2) Frequent Pattern. Emerging pattern means whose frequency change significantly from one dataset to another. Frequent pattern-Tree use generate and test approach, it generates candidate itemset and then test if they are frequent. In this paper also describe the FP-Growth and important of correlation analysis between patterns. Gene expression data means the symptoms of living beings. EP and FP are apply on gene data and reduce the data of the gene dataset.**

***Index Terms*- EP, FP, Apriori algorithm, correlation, gene data**

I. INTRODUCTION

A gene is the molecular unit of heredity of a living organism. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. All organisms have genes corresponding to various biological traits, some of which are instantly visible.

A FP-Tree is generate the candidate itemset. And A-Priori algorithm is reduce the candidate itemsets. And now we apply mostly FP-Growth algorithm which is not generate candidate itemset. It is also called without candidate generation algorithm.

Emerging Pattern represent the frequently changes the data. It represents strong contrast knowledge. It is new type of knowledge pattern which compare the two classes of data. EP is an itemset whose support in one data differs from its support in another [2].

II. LITERATURE REVIEW

Frequent Pattern Tree Algorithm:

From some time ago we use FP-Tree algorithm. Market basket analysis is just one form of frequent pattern mining. In fact, there are many kinds of frequent patterns, association rules, and correlation relationships.

A FP-Tree is an extended prefix tree structure that represents the transaction database in a compact and complete way. Only frequent length-1 items will have nodes in the tree, and the tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of sharing nodes than less frequently occurring ones. Each transaction in the database is mapped to one path in the FP-Tree. The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database [13].

Disadvantage: FP-Tree generate the candidate itemsets which take more time to given output and occupies more space.

Apriori Algo

- Main Steps of Apriori Algorithm:
  - Use frequent (k – 1)-itemsets (L_{k-1}) to generate candidates of frequent k-itemsets C_k
  - Scan database and count each pattern in C_k , get frequent k-itemsets (L_k)
- We find the support and with the help of support we generate candidate itemsets.

$$Support = P\,(A \cup B)$$ [1]

Below the example of apriori algo which min_supp = 2.

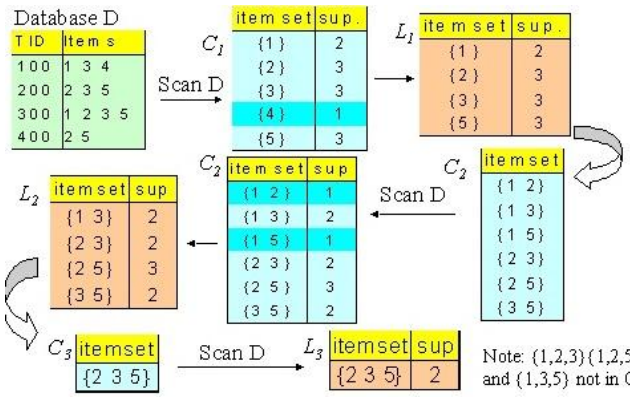

Fig 1: Apriori Algo Example[15]

Apriori Algorithm also generate candidate itemset but it generate less itemset so it is better compare to FP-tree.

Frequent Pattern Growth Structure (FP-Growth)

It allows frequent itemset discovery without candidate itemset generation. Two step approach:

- Step 1 : Build a compact data structure called the FP-tree

- Built using 2 passes over the data-set.

- Step 2 : Extracts frequent itemsets directly from the FP-tree
  - Traversal through FP-Tree

Emerging Pattern

Given two or more datasets contrast patterns are patterns that describe significant differences between the given datasets. A pattern is considered as describing differences between the two datasets if some statistics (e.g., support or risk ratio) for with respect to each of the datasets are highly different.

We often refer to the dataset/class where a pattern has the highest frequency as its home dataset/class. Many names have been used to describe contrast patterns, including emerging patterns, contrast sets, group differences, patterns characterizing change, classification rules and discriminating patterns [2]. Contrast data mining can also be applied to many types of data, including vector data, transaction data, sequence data, graph data, image data and data cubes [2]. Emerging Patterns are those whose frequencies change significantly from one dataset to another. They represent strong contrast knowledge. It is the new type of knowledge pattern that describes significant changes (differences or trends) between two classes of data. An Emerging Pattern is an itemset whose support in one set of data differs from its support in another [2]. Following are the various types of Emerging Patterns which are proposed till now:

A. ρ- Emerging Patterns (ρ- EP)

Given two different classes of datasets D1 and D2, the growth rate of an itemset X from D1 to D2 is defined as

$$GR(x) = \begin{cases} 0, if\ \sup_1(x) = 0 and \sup_2(x) = 0 \\ \infty, if\ \sup_1(x) = 0 and \sup_2(x) > 0 \\ \dfrac{\sup_2(x)}{\sup_1(x)}, otherwise \end{cases} \quad (1.1)$$

Emerging Patterns are those itemsets with large growth rates from D1 to D2. Given a growth rate threshold ρ > 1, an itemset X is said to be a ρ - Emerging Pattern (ρ-EP or simply EP) from a background dataset D1 to a target dataset D2 if GR (X)> =ρ. [4]

B. Jumping Emerging Patterns (JEP)

The strength of an EP X is defined as

$$strength(x) = \frac{GR(x)}{GR(x)+1} * \sup(x) \quad (1.2)$$

A Jumping Emerging Pattern (JEP) from a background dataset D1 to a target dataset D2 is defined as an Emerging Pattern from D1 to D2 with the growth rate of . Note that for a JEP X, strength(X) = supp(X). [4]

C. Essential Jumping Emerging Patterns (EJEP)

EJEPs are defined as minimal itemsets whose supports in one data class are zero but in another are above a given support threshold ξ. Given ξ > 0 as a minimum support threshold, an Essential Jumping Emerging Pattern (EJEP) from D1 to D2, is an itemset X that satisfies the following conditions: [4]

1. $\text{supp}_{D1}(X) = 0$ and $\text{supp}_{D2}(X) > \xi$, and

2. Any proper subset of X does not satisfy condition 1.

TABLE I
COMPARISON BETWEEN JEP AND EJEP

Type	supp(D1)	supp(D2)	GR	Minimal
JEP	0	> 0	1	NO
EJEP	0	$> \mu$	1	YES

D. Chi- Emerging Patterns (Chi-EP)

We say that an itemset, X, is a Chi Emerging Pattern (Chi EP), if all the following conditions about X are true: [4]

1. Supp(x) >= ξ, where ξ is a minimum support threshold;

2. GR(x) >= ρ, where ρ is a minimum growth rate threshold;

3. It has larger growth rate than its subsets;

4. It is highly correlated according to common statistical measures such as chi-square value. Length-1 itemsets, that satisfy the above three conditions, pass chi-square test directly.

E. Noise Tolerant Emerging Patterns (NEP)

According to different types of the training data, the strategies of the EPs can be divided into two categories, i.e., the EPs with the infinite growth rate and the EPs with the finite growth rate. [6]

The EJEP strategy only cares about those itemsets with the infinite growth rate. It ignores those patterns which have very large growth rates, although not infinite, i.e., the so called "noise". However, the real-world data always contains noises and the NEP strategy considers noises and provides higher accuracy than the EJEP strategy. [6]

EJEPs allow noise tolerance in dataset D2. However, real-world data always contains noises in both dataset D1 and dataset D2. Both JEPs and EJEPs cannot capture those useful patterns whose support in dataset D1 is very small but not strictly zero; that is, they appear only several times due to random noises. Therefore the Noise-tolerant EPs were proposed. [6]

F. High Growth-Rate Emerging Patterns (HGEP)

Although the NEP strategy takes noise patterns into consideration, it still will miss some itemsets with a large growth rate, which may result in the low accuracy. Therefore High Growth-rate EP (HGEP) strategy was proposed to improve the disadvantage of the NEP strategy. [6] High Growth-Rate Emerging Pattern (HGEP), which can improve the accuracy of a classifier. [6] An itemset X is an HGEP for dataset D2 from dataset D1 to dataset D2, if X satisfies one of the following two conditions: where δ1 and δ2 are the support thresholds of the dataset D1 and D2.

Condition 1:

1.1 $0 < supp_{D1}(X) \leqq \delta 1$ and $supp_{D2}(X) \geqq \delta 2$, Where $\delta 1 << \delta 2$.

1.2 GR (propersubset(X)) < GR(X).

Condition 2:

2.1 $supp_{D1}(X) = 0$ and $supp_{D2}(X) >= \delta 2$.

2.2 Any proper subset of X does not satisfy Condition.

They have the following properties: [6]

EP ⊇ JEP ⊇ EJEP

NEP ⊇ EJEP and HGEP ⊇ EJEP

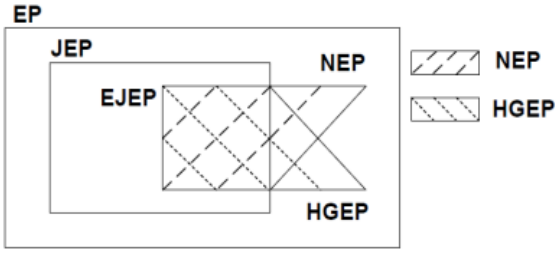

Fig. 2 : The relationships between various EPs.

III. FREQUENT PATTERN TREE STRUCTURE BASED (FP- GROWTH) ALGORITHM

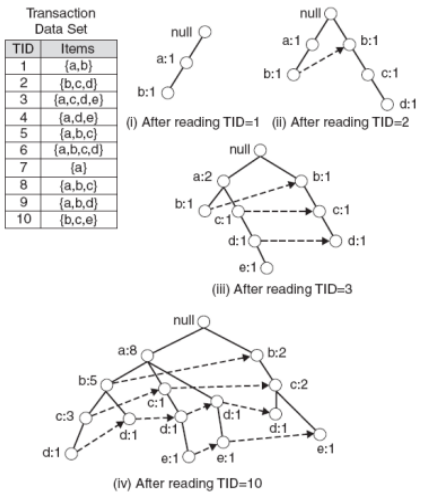

Fig 3 : Generate FP Tree [12]

- Nodes correspond to items and have a counter
- FP-Growth reads 1 transaction at a time and maps it to a path
- Fixed order is used, so paths can overlap when transactions share items (when they have the same prex).
- In this case, counters are incremented
- Pointers are maintained between no discontinuing the same item, creating singly linked lists (dotted lines)
- The more paths that overlap, the higher the compression. FP-tree may fit in memory.
- Frequent itemsets extracted from the FP-Tree

TABLE II
FP-TREE V/S FP-GROWTH

Parameters	**FP Tree (Apriori Algo)**	**FP-Growth**
Techniques	Use Apriori property join and prune property	It constructs conditional FP Tree & conditional pattern base from database which satisfy minimum support
Memory Utilization	Due to large number of candidate are generated so require large	Due to compact structure and number of candidate generation require less memory
No. Of Scans	Multiple scans for generating candidate sets	Scan the DB only twice and twice only
Time	Execution time is more as time is wasted in producing candidate every	Execution time is small than Apriori Algorithm

Correlation Analysis [3]:

For a given instance, the distinct values of a given explanatory variable can pull up (higher value) or pull down (lower value) the model output. The proposed idea is to analyze the relationship between the values of an input variable and the probability of occurrence of a given target class. The goal is to increase (or decrease) the score, the target class probability, by exploring the different values taken by the explanatory variable. For instance for medical data one tries to decrease the probability of a disease; in case of cross-selling one tries to increase the appetency to a product; and in government data cases one tries to define a policy to reach specific goals in terms of specific indicators (for example decrease the unemployment rate).

IV. RELATED WORK

When original gene expression data are used then some problems are occurred.

• The large search space:

A microarray gene expression database consists of the data obtained from many microarray slides under various experimental conditions. Each microarray slide can be considered as one database transaction containing the values of genes in one experimental condition, and each gene can be considered as one data item. For human beings there are 50,000–100,000 genes. There would be a tremendous number of candidate itemsets that must be identified by an association rule mining algorithm. For such an algorithm to work effectively, it must be able to deal robustly with the dimensionality of this feature space

• Uninteresting genes:

Not all genes are interesting to biologists. Sometimes biologists may be interested in some special genes. So they may just want to mine the association rules among these interesting genes and do not want to waste time to mine all other genes_ possible association rules.

We use these steps and reduce the gene dataset and easily and speedy get output.

Step 1: Read the gene data file and gives number of rows and columns.

Step 2: Split the numbered data into two datasets
Positive – tumor biopsies
Negative – normal biopsies

Step 3: Apply FP growth algorithm into both datasets.

Step 4: On these outputs, given min-ratio and find Emerging Pattern.

Step 5: Convert numbered EP into gene dataset.

Step 6: Find highly correlated genes.

V. OPEN CHALLENGES

A. Find Information Gain from large Datasets.

B. Find more interesting patterns from large microarray dataset.

VI. CONCLUSION

This paper provides study of Emerging Patterns in the field of data mining and Knowledge Discovery in Databases (KDD). Specifically, it has investigated the following problems: (1) how to define various kinds of Emerging Patterns that provide insightful knowledge and are useful for classification; (2) how to mine those useful Emerging Patterns.

Based on the comparison with the FP-Tree and FP-Growth, FP-Growth is occupied less memory and it is taken to decrease the time of output. And find correlation between EP and gave accurate result.

REFERENCES

[1] Jiawei Han, Micheline Kamber; Data Mining: Concepts and Techniques; 2nd ed.; Morgan Kaufmann Publishers, 2006

[2] Kotagiri Ramamohanarao, James Bailey and Hongjian Fan, "Efficient Mining of Contrast Patterns and Their Applications to Classification", IEEE Society, ICISIP 2005, pp. 39-47

[3] Vincent Lemair, Carine Hue, Olivier Bernier, "Correlation Analysis in Classifiers", IEEE Society, 2011

[4] Hongjian FAN, "Efficient Mining of Interesting Emerging Patterns and Their Effective Use in Classification", The University of Melbourne, 2004

[5] Heikki Mannila, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Kluwer Academic Publishers, 2004

[6] Ye-In Chang, Zih-Siang Chen, and Tsung-Bin Yang ,"A High Growth-Rate Emerging Pattern for Data Classification in Microarray Databases", Lecture Notes on Information Theory Vol. 1, No. 1, March 2013

[7] Kotagiri Ramamohanarao, Thomas Manoukian and James Bailey, "Fast Algorithms for Mining Emerging Patterns", Springer 2002

[8] Kotagiri Ramamohanarao and James Bailey, "Discovery of Emerging Patterns and Their Use in Classification", Springer, 2003

[9] Liang Wang, Yizhou Wang and Debin Zhao, "Building Emerging Pattern (EP) Random Forest for Recognization",IEEE Society, ISIP 2010

[10] Kotagiri Ramamohanarao, Qun Sun and Xiuzhen Zhang, "Noise Tolerance of EP Based Classifiers", Springer 2003

[11] FP Tree,

https://fenix.tecnico.ulisboa.pt/downloadFile/377957125009

6/licao_10.pdf

[12] FP Growth,

http://www.florian.verhein.com/teaching/2008-01-09/fp-

growth-presentation_v1%20(handout).pdf

[13] FP Tree Structure,
http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm

[14] FP Growth Algo,

http://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf

[15] Figure Of Apriori,
http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput499/slides/Lect10/sld054.htm