ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 10, S.Iss 2, 2021

Literature Review: Emerging Patterns and Frequent Pattern Growth Algorithm Applied to Gene Expression Data

[1]Shail Dubey, [2]Ashish Shukla, [3]Shalini Gupta, [4]Abhay Shukla, [5]Rituraj Kushwaha, [6]Pooja Diwivedi

Department of Computer Applications, Axis Institute of Higher Education, Kanpur, Uttar Pradesh, India

Abstract

Data mining involves extracting Knowledge Discovery in Databases (KDD) is a comprehensive process of extracting useful and previously unknown information from large data sets. It discovers intriguing or valuable patterns and connections within the data.Two types of patterns are discussed: (1) Emerging Patterns and (2) Frequent Patterns. Emerging Patterns are those whose frequency changes significantly between datasets. The Frequent Pattern-Tree method uses a generate-and-test approach, where candidate item sets are generated and then tested for frequency. This paper also covers the FP-Growth algorithm and emphasizes the significance of correlation analysis between patterns. Gene expression data refers to the characteristics of living organisms. Emerging Patterns and Frequent Patterns are applied to gene data to reduce the gene dataset.

I.INTRODUCTION

A gene, s the essential units of heredity in living organisms. Genes are essential as they dictate the production of all proteins and functional RNA chains in an organism. They contain the instructions necessary to build and sustain the cells of an organism and Passing Genetic Traits to Offspring. The Role of Genes in All Organisms possess genes that correspond Genes and the Inheritance of Biological Traits, some of which are readily observable.

A Frequent Pattern sapling is used to cause applicant item sets, while the Apriori algorithm reduces these applicant item sets. Nowadays, the FP-Growth algorithm is commonly applied, which does not generate candidate item sets. This is why it is also known as the algorithm without candidate generation.

Emerging Patterns (EPs) represent data that changes frequently and highlights strong contrast knowledge. This novel type of knowledge pattern contrasts two categories of data. The Emerging Pattern (EP) is Understanding Item Support in Data Mining.

II.LITERATUREREVIEW

Association rule mining is a technique in data mining used to discover interesting relationships and patterns among items in large datasets. The foundational work in association rule mining was laid by Agrawal et al. (1993), who introduced the Apriori algorithm, which identifies frequent item sets and generates association rules from these item sets.

The Apriori algorithm works by iteratively identifying frequent item sets and pruning infrequent ones. It uses a breadth-first search approach and a candidate generation method to find item sets that meet a minimum support threshold. The algorithm's primary strength is its

ability to handle large datasets by focusing only on those item sets that have sufficient support in the context of data mining, particularly in association rule mining and similar analyses, each transaction in a database is mapped to identify patterns, item sets, or rules. This mapping process is crucial for extracting meaningful insights and discovering associations among items within transactions.13】.

Disadvantage: The Frequent Pattern generates applicant item sets, which require extra time to produce result and occupy extra space.

Apriori Algo.

Generate Candidate Item Sets:

Use frequent (k−1) (k-1) (k−1)-item sets (Lk−1L_{k-1} Lk−1) to generate candidates of frequent kkk-item sets (CkC_kCk).

Scan Database:

Scan the database and count the occurrences of each pattern in CkC_kCk.

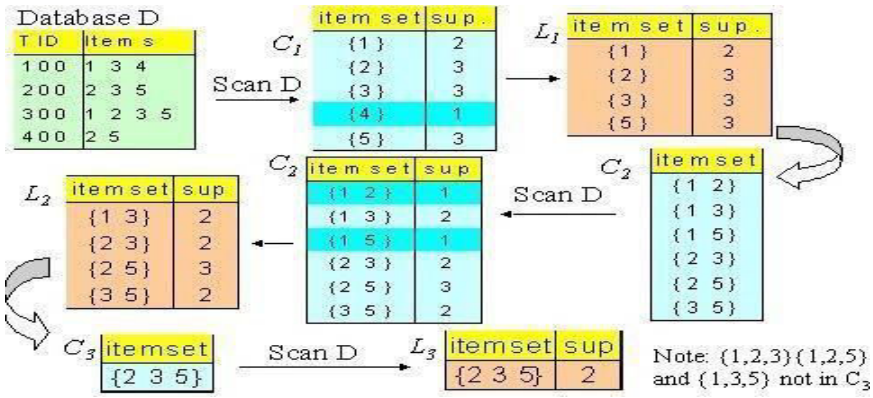
Identify the frequent kkk-item sets (LkL_kLk) by selecting candidates that meet the minimum support threshold.

Generate Candidate Item Sets: Calculate the support for each candidate item set.

Use the support values to generate new candidate item sets for the next iteration.

The process is repeated until no more frequent item sets can be found.

$Support$=(AUB)[1]



Example of the apriori algorithm with minimum support(min_supp)=2.

Fig1: Example Apriori Algo

Frequent Pattern Growth Structure (FP-Growth)

The Apriori Algorithm also generates candidate item sets, but it produces fewer item sets compared to the FP-Tree. This makes it more efficient and preferable in certain scenarios, as it requires less memory and computational time.

The FP-Growth algorithm is an efficient approach for mining frequent item sets without the need for candidate itemset generation, which is a key limitation of the Apriori algorithm. It uses a two-step approach involving the construction of a compact data structure called the FP-tree (Frequent Pattern Tree) and the mining of frequent item sets from this structure:

Step1: Building the FP-Tree: A Step -by- Step Guide Built using 2 passes over the data-set.

Step 2: Extracts frequent item sets directly from the FP-tree Travers although FP-Tree.

Emerging Pattern

Emerging patterns are patterns that show a significant increase in frequency or significance in

a dataset compared to a previous period or different contexts. They are used to identify trends or shifts in data that may not be apparent from static analysis. Emerging patterns are particularly useful for discovering new and evolving trends in various domains, such as finance, healthcare, and marketing. These are patterns that become more frequent or significant over time or across different contexts. They are often used to identify novel trends, anomalies, or shifts in the data.

Emerging Patterns are those whose frequencies change significantly between datasets, representing strong contrast knowledge. This new type of knowledge pattern describes outstanding replace (differences or trends) different types of classes of data An Emerging Pattern is an item set whose carry varies significantly between two datasets.. Below are the different types of Emerging Patterns proposed to date:

A.ρ-Emerging Patterns(ρ-EP)

The growth rate of an item set X from dataset D1 to dataset D2 quantifies how the frequency or significance of X changes between these two dataset as Growth Rate=Support(X in D2)−Support(X in D1)Support(X in D1)\text{Growth Rate} = \franc{\text{Support}(X \text{ in } D2) - \text{Support}(X \text{ in } D1)}{\text{Support}(X \text{ in } D1)}Growth Rate=Support(X in D1)Support(X in D2)−Support(X in D1)

where Support (X in D1)\text{Support}(X \text{ in } D1)Support(X in D1) and Support(X in D2)\text{Support}(X \text{ in } D2)Support(X in D2) represent the support levels of the itemset X in datasets D1 and D2, respectively.

□0,ifsup(x)□0andsup(x)□0

□ 1 2

Identifying Emerging Patterns with Growth Rate from D1 to D2

is defined by its growth rate. For a Jumping Emerging Pattern (JEP) X, the strength of X is given by its support, strength(X)=supp(X)\text{strength}(X) = \text{supp}(X)strength(X)=supp(X)【4】.

Essential Jumping Emerging Patterns (EJEP)

Necessary Jumping Emerging Patterns (EJEPs) are minimum item sets with zero support in one data class but support above a given threshold ξ\xiξ in another data class. Formally, an EJEP from dataset D1 to dataset D2 is an item set X that meets the following conditions given a minimum support threshold θ\thetaθ and a minimum growth rate ρ\rhoρ:

Minimum Support in D2: The item set X must have a support in the target dataset D2 that is greater than or equivalent the minimum support holdθ\thetaθ.

Support (X in D2)≥θ\text{Support}(X \text{ in } D2) \geq \theta Support(X in D2)≥θ

Jumping Condition: The item set X should not appear in the background dataset D1 or have a support close to zero in D1.

Support (X in D1)≈0\text{Support}(X \text{ in } D1) \approx 0Support(X in D1)≈0

Growth Rate: The growth rate of X from D1 to D2 should be equal to or greater than the minimum growth rate ρ\rhoρ.

Growth Rate(X)=Support (X in D2) Support (X in D1)+ϵ≥ρ\text{ Growth Rate}(X) = \frac{\text{Support}(X \text{ in } D2)}{\text{Support}(X \text{ in } D1) + \epsilon} \geq \rhoGrowth Rate(X)=Support(X in D1)+ϵSupport(X in D2)≥ρ

Here, ϵ\epsilonϵ is a small constant added to avoid division by zero.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 10, S.Iss 2, 2021

Essential Condition: The item set X should be a minimal pattern that satisfies the above conditions, meaning that no proper subset of X meets these conditions.

By adhering to these conditions, an Essential Jumping Emerging Pattern (EJEP) captures item sets that exhibit a significant increase in support from D1 to D2, providing insightful knowledge for classification and other data mining tasks.

support threshold ξ>0\xi > 0ξ>0:

The support of X in dataset D1 is zero.

The support of X in dataset D2 is above the threshold ξ\xiξ.

supp(D1,X): The support count or frequency of itemset X in dataset D1.

supp(D2,X)\text{supp}(D2, X)supp(D2,X): The support count or frequency of itemset X in dataset D2

Any proper subset of X not fulfill thecondition1.

TABLE I: COMPARISON BETWEEN JEPANDEJEP

Type	supp(D1)	supp(D2)	GR	Minimal
JEP	0	>0	1	NO
EJEP	0	> □	1	YES

Chi-Emerging Patterns (Chi-EP)

Formally, an Essential Jumping Emerging Pattern (EJEP) from dataset D1 to dataset D2 is an item set X that meets the following conditions given a minimum support threshold θ\thetaθ and a minimum growth rate ρ\rhoρ【4】:

Minimum Support Threshold:

The support of X is greater than or equal to a minimum support threshold ξ\xiξ, i.e., supp(X)≥ξ\text{supp}(X) \geq \xisupp(X)≥ξ.

Minimum Growth Rate Threshold:

An item set X is considered a Chi-Emerging Pattern (Chi-EP) if it satisfies all of the following conditions:

Support in D2: The item setX must have a support in the target dataset D2 that is greater than a specified threshold ξ\xiξ.

SupportD2(X)>ξ\text{Support}_{D2}(X) > \xiSupportD2(X)>ξ

Chi-Square Test: The item set X must show a statistically significant association with the target dataset D2 as determined by the chi-square test. This can be formalized as:

χ2(X)≥χα2\chi^2(X) \geq \chi^2_{\alpha}χ2(X)≥χα2

whereχ2(X)\chi^2(X)χ2(X) is the chi-square statistic for the item set X, and χα2\chi^2_{\alpha}χα2 is the critical value of the chi-square distribution for a given significance level α\alphaα.

Growth Rate: The item set X should have a significant growth rate from the background dataset D1 to the target dataset D2. Specifically, the growth rate of X is greater than or equal to a minimum growth rate threshold ρ\rhoρ:

GR(X)≥ρ\text{GR}(X) \geq \rhoGR(X)≥ρ

The growth rate GR(X)\text{GR}(X)GR(X) is calculated as:

GR(X)=SupportD2(X)SupportD1(X)+ϵ\text{GR}(X) = \frac{\text{Support}_{D2}(X)}{\text{Support}_{D1}(X) + \epsilon}GR(X)=SupportD1(X)+ϵSupportD2(X)

whereϵ\epsilonϵ is a small constant added to avoid division by zero.

Support in D1: The support of the item set X in the background dataset D1 should be relatively low or zero.

SupportD1(X)≈0\text{Support}_{D1}(X) \approx 0SupportD1(X)≈0.

Minimality Condition: The item set X should be a minimal pattern that satisfies the above conditions, meaning that no proper subset of X meets these conditions.

By meeting these conditions, a Chi-Emerging Pattern (Chi-EP) not only indicates a significant increase in support from D1 to D2 but also ensures that this increase is statistically significant and substantial, providing robust and insightful patterns for classification and other data mining tasks.

Larger Growth Rate than Subsets: X has a larger growth rate compared to all its subsets.

High Correlation: The item set X is considered a Chi-Emerging Pattern (Chi-EP) if it satisfies all of the following conditions:

Support in D2: The item set X must have a support in the target dataset D2 that is greater than a specified threshold ξ\xiξ.

SupportD2(X)>ξ\text{Support}_{D2}(X) > \xiSupportD2(X)>ξ

Chi-Square Test: The item set X must show a statistically significant association with the target dataset D2 according to the chi-square test. For length-1 item sets, the chi-square test is applied directly:

χ2(X)≥χα2\chi^2(X) \geq \chi^2_{\alpha}χ2(X)≥χα2

whereχ2(X)\chi^2(X)χ2(X) is the chi-square statistic for the item set X, and χα2\chi^2_{\alpha}χα2 is the critical value of

GR(x)□□□,ifsup1(x)□0andsup2(x)□0sup(x)(1.1)

2, otherwise □□sup1(x)

EP (Emerging Patterns) sets of items that exhibit significant growth from dataset D1 to dataset D2. If we have a growth rate threshold ρ greater than 1, an item set X is classified as a ρ-Emerging Pattern (ρ-EP, or simply EP) when its growth rate (GR) When evaluating Emerging Patterns dataset D1 to the target dataset D2 meets or exceeds the threshold ρ.

B. Jumping Emerging Patterns (JEP)

The power of an EPX is defined as

GR(x)When dealing with different types of training data, strategies for identifying Emerging Patterns (EPs) can be divided into two categories:

EPs with Infinite Growth Rate:

Definition: These are itemsets whose support in the background dataset D1 is zero and whose support in the target dataset D2 is greater than a specified threshold.

Conditions:

SupportD1(X)=0andSupportD2(X)>ξ\text{Support}_{D1}(X) = 0 \quad \text{and} \quad \text{Support}_{D2}(X) > \xiSupportD1(X)=0andSupportD2(X)>ξ

Example: Consider an item set X that does not appear in the background dataset (support is zero) but appears frequently in These strategies help in effectively categorizing and analyzing

emerging patterns based on their growth characteristics, enabling better insights for classification and data mining tasks.

The Essential Jumping Emerging Pattern (EJEP) strategy focuses specifically on itemsets that exhibit significant changes between two datasetsinfinite growth rate. It ignores those patterns

The NEP (Noise-tolerant Emerging Patterns) strategy handles data with very large growth rates, often referred to as "noise." Despite not being infinite, this noise is inherent in real-world data.

The NEP strategy effectively considers noise and provides the chi-square distribution for a given significance level α\alphaα.

Growth Rate: The item set X should have a significant growth rate from the background dataset D1 to the target dataset D2. The growth rate of X is greater than or equal to a minimum growth rate threshold ρ\rhoρ:

GR(X)≥ρ\text{GR}(X) \geq \rhoGR(X)≥ρ

The growth rate GR(X)\text{GR}(X)GR(X) is calculated as:

GR(X)=SupportD2(X)SupportD1(X)+ϵ\text{GR}(X) = \frac{\text{Support}_{D2}(X)}{\text{Support}_{D1}(X) + \epsilon}GR(X)=SupportD1(X)+ϵSupportD2(X)

whereϵ\epsilonϵ is a small constant added to avoid division by zero.

Support in D1: The support of the item set X in the background dataset D1 should be relatively low or zero.

SupportD1(X)≈0\text{Support}_{D1}(X) \approx 0SupportD1(X)≈0

Minimality Condition: The item set X should be a minimal pattern that satisfies the above conditions, meaning that no proper subset of X meets these conditions.

By satisfying these conditions, a Chi-Emerging Pattern (Chi-EP) indicates a significant increase in support from D1 to D2, ensuring that this increase is statistically significant and substantial. This makes Chi-EPs valuable for identifying robust and insightful patterns for classification and other data mining tasks.

Noise Tolerant Emerging Patterns (NEP)

the target dataset. Since the support in D1 is zero, the growth rate is effectively infinite.

Usefulness: EPs with infinite growth rates are particularly useful for identifying completely new patterns or trends that emerge in the target dataset but are absent in the background dataset.

EPs with Finite Growth Rate:

Definition: These are item sets whose support in the background dataset D1 is greater than zero and whose growth rate from D1 to D2 meets or exceeds a specified threshold ρ\rhoρ.

Conditions: SupportD1(X)>0andGR(X)≥ρ\text{Support}_{D1}(X) > 0 \quad \text{and} \quad \text{GR}(X) \geq \rhoSupportD1(X)>0andGR(X)≥ρ

Growth Rate Calculation: GR(X)=SupportD2(X)SupportD1(X)\text{GR}(X) = \frac{\text{Support}_{D2}(X)}{\text{Support}_{D1}(X)}GR(X)=SupportD1(X)SupportD2(X)

Example: Consider an item set X that has some support in both datasets, but its support in D2 is significantly higher than in D1. The finite growth rate quantifies this increase.

Usefulness: EPs with finite growth rates are useful for identifying patterns that are already

present in the background dataset but become significantly more prevalent in the target dataset.

higher accuracy compared to the EJEP (Essential Jumping Emerging Patterns) strategy.

EJEPs, while accommodating noise tolerance in dataset D2, may not be as effective when noise is present in both datasets D1 and D2. Both JEPs (Jumping Emerging Patterns) and EJEPs struggle to manage this noise effective.

To capture useful patterns whose support in the background dataset D1 is very small but not strictly Zero Noise-Tolerant Emerging Patterns are proposed. These pattems are designed to identify significant changes trends that may be obscured by noise in the data.

strength(x)□ □GR(x)□1*sup(x) (1.2)

.F.High Growth-Rate Emerging Patterns(HGEP)

The High Growth-Rate Emerging Patterns (HGEP) strategy was designed to identify item sets that exhibit exceptionally high growth rates from one dataset to another. This strategy focuses on detecting patterns that show significant increases in support or importance, indicating a notable change or trend main goal of HGEP is to pinpoint item sets that have experienced a dramatic increase in their support from the background dataset D1D1D1 to the target dataset D2D2D2. This helps in identifying patterns that have become significantly more relevant or prevalent.

Absolute Growth Rate: Calculate the absolute difference in support between D2D2D2 and D1D1D1: Absolute Growth Rate(X)=supp(D2,X)−supp(D1,X)\text{Absolute Growth Rate}(X) = \text{supp}(D2, X) - \text{supp}(D1, X)Absolute Growth Rate(X)=supp(D2,X)−supp(D1,X)

Relative Growth Rate: Alternatively, compute the relative growth rate as a proportion of the initial support:

Relative Growth Rate(X)=supp(D2,X)−supp(D1,X)supp(D1,X)\ Condition2:

suppD1(X)=0andsuppD2(X)>=δ2.

Any proper subset of X does not satisfy.

The properties of HGEP are following as: [6]

1EP⊇JEP⊇EJEP

NEP⊇EJEPandHGEP⊇EJEP

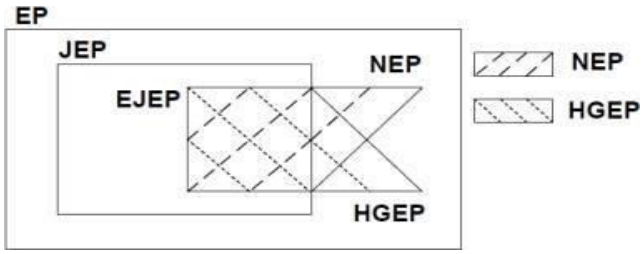


Fig. 2: The relationships between various EPs.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 10, S.Iss 2, 2021

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

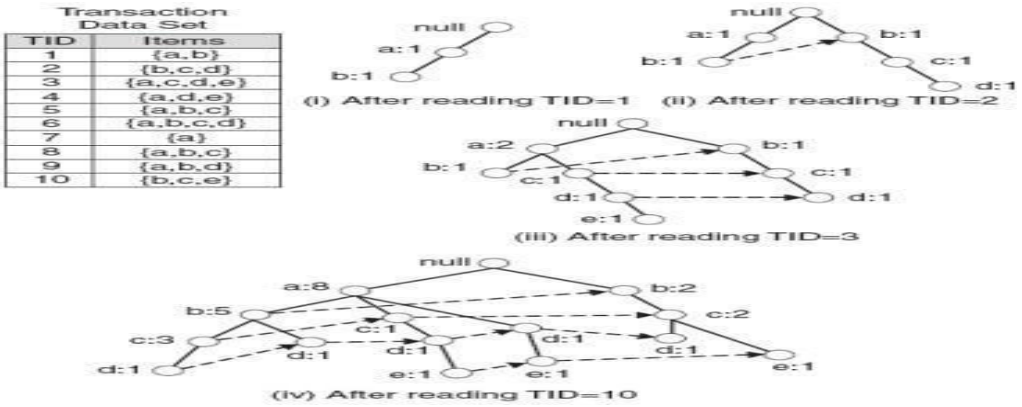


Fig3: Generate FPTree [12]

Nodes correspond to items and have a counter

FP-Grow threads 1 transaction at a time and maps path Correlation Analysis[3]:

As an independent or predictor variable either raise or lower the model's output. The proposed approach involves analyzing how different Values of an input variable impact)\text{Relative Growth Rate}(X) = \frac{\text{supp}(D2, X) - \text{supp}(D1, X)}{\text{supp}(D1, X)}Relative Growth Rate(X)=supp(D1,X)supp(D2,X)−supp(D1,X)

 Fixed Order Usage: Transactions or sequences are processed considering the exact order in which items appear. This means that the sequence {A, B, C} is treated as distinct from {C, B, A}, and the order of items is critical for pattern discovery.

Increment Counters: In cases of overlapping paths, counters for the shared items are incremented.

Pointer Maintenance: Pointers are maintained to avoid discontinuity in the same item, resulting in singly linked lists).

Higher Compression: Greater path overlap results in higher compression, allowing the FP-Tree to fit in memory more effectively.

Frequent Item Sets: Frequent item sets are extracted from the FP-Tree based on the stored information.

the probability of a specific target class. The objective is to adjust the target class cases by examining different values of the explanatory variable. For example:

In medical information, the objective might be to remove the cases of a disease.

In selling scenarios, the aim could be to increase the likelihood of a customer purchasing a product.

In government data analysis, the focus might be on crafting policies to achieve specific goals, such as reducing the unemployment rate.

IV. RELATED WORK

When using original gene expression data, several challenges arise:

Large search space: A mini array gene expression database contains data from numerous microarray slides under various experimental conditions. Each slide can be viewed as a single database transaction that holds gene values for one experimental condition, with each gene

representing a data item. For humans, there are about 50,000 to 100,000 genes, leading to an enormous number of candidate item sets that an association rule mining algorithm must identify. For the algorithm to be effective, it must handle the high dimensionality of this feature space robustly.

Uninteresting genes: Not all genes are of interest to biologists. Sometimes, biologists focus on specific genes and want to mine association rules only among these genes, avoiding the time-consuming process of mining all possible association rules among all genes.

To address these issues and streamline the gene dataset for quicker and easier analysis, the following steps are used:

Study the gene info. file and determine the no. of rows
and columns.

Split the numbered data into two datasets: Positive (tumor biopsies) and Negative (normal biopsies).

Apply the FP-growth algorithm to both datasets.

Determine the Emerging Patterns (EPs) by applying a minimum ratio to the outputs.

Convert the numbered EPs back into the gene dataset.

Identify highly correlated genes.

OPEN CHALLENGES

o find Information Gain from large datasets and discover interesting patterns in large microarray datasets, you can follow a systematic approach. Here's a detailed guide on how to achieve both tasks:

1. Finding Information Gain

Information Gain is a measure used to determine the effectiveness of a feature in separating different classes in a dataset. It is commonly used in decision tree algorithms. Here's how you can calculate Information Gain for large datasets using Python:

Load the Data:

Load your dataset into a pandas Data Frame.

Calculate Entropy and Information Gain:

Define functions to calculate entropy and information gain.

CONCLUSION

This paper explores the study of Emerging Patterns within the field of data mining and Knowledge Discovery in Databases (KDD). Specifically, it addresses two key problems:

Defining Emerging Patterns offer valuable insights by highlighting significant changes and trends in data. By leveraging these patterns, you can enhance classification models, uncover meaningful features, and gain a deeper understanding of the underlying data.

Mining those useful Emerging Patterns.

A comparative analysis of FP-Tree and FP-Growth shows that FP-Growth occupies less memory and is more efficient.

To reduce the time of result and detect the correlation between Emerging Patterns (EPs) for accurate results, the following steps can be taken:

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 10, S.Iss 2, 2021

REFERENCES

1. Jiawei Han, Micheline Kamber; Data Mining: Concepts and Techniques;2nded.; Morgan KaufmannPublishers,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 UniversityofMelbourne,2004
5. Heikki Mannila, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", KluwerAcademicPublishers,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 TheoryVol.1, No. 1, March2013
7. Kotagiri Ramamohanarao, Thomas Manoukian and James Bailey, "Fast Algorithms for Mining Emerging Patterns", Springer2002
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 Recognition", IEEE Society, ISIP2010
10. Kotagiri Ramamohanarao, Qun Sunand Xiuzhen Zhang, "Noise Tolerance of EP Based Classifiers", Springer2003