Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

MISDSSP-Tree: A Novel Minimum Item Support Difference Based Pattern Tree Mining Approach For Mining Rare And Frequent Patterns

[1] Keerti Shrivastava , [2] Dr. Varsha Jotwani

[1] Rabindranath Tagore University, Bhopal.

[2] Rabindranath Tagore University, Bhopal.

Abstract— Life-threatening diseases are a major worry for many people in today's society. To limit the severity of the side effects of diseases, early identification and diagnosis are crucial. Rare Association Rule Mining (RARM), a method of computational intelligence, may be widely used in the study of illness. RARM study is dependent upon the assumption that all data to be mined is immediately accessible at beginning of the mining process. With the addition of new entries or deletion of old ones, medical databases in use might change over time. In addition, when the database is updated, the user may choose a new threshold for creating an appropriate collection of unusual association rules. A straightforward, but inept method may become the reconstruction of the whole mining algorithm from scratch, with each updated dataset and the revised threshold level for the present set of rare association rules. This paper presents an effective method in the context of rare patterns for 3 adverse diseases: hepatitis, breast cancer, & cardiovascular, for identification of symptoms & risky factors as described in the rare association rules. The Minimum Item Support Difference Single Scan Pattern-Tree (MISDSSP-tree) algorithm is used to calculate the support difference of each itemset. It checks the support count difference for each item with minimum support difference and compares it to the MIS value. MIS value satisfied itemset comprise frequent items & rare items. Experimental outcomes on real-life data sets demonstrated that presented MISDSSP-tree-algorithm increases performance against previous SSP-tree and ISSP-tree approaches by reducing the explosion of frequent itemset that contains frequent patterns and often comprise rare items. The MISDSSP-tree algorithm outperforms runtime and memory consumption. The relevance of the proposed approach over the usual strategy of repeatedly mining the complete updated database is shown via experimental analysis.

Keywords— Data Mining, Association Rule Mining, Frequent and Rare Pattern Mining, Adverse Diseases, MIS, Support Difference, SSP-Tree.

I. INTRODUCTION

Today many researchers are using data mining techniques for knowledge discovery from massive data. Data is being gathered & accumulated at a rapid rate in all fields. Computational theories & techniques are urgently needed to aid people in extracting usable information (knowledge) from fast-increasing amounts of digital data. Deployment of particular DM (Data Mining) techniques for the identification and extraction of patterns

is at the core of the procedure (Kumar, Jain and Chauhan, 2019). Characterization, generalization, clustering, classification, evolution, association, data visualization, pattern matching, & meta-rule guided mining are just a few of the data mining approaches that have recently been created (Sumathi and Sivanandam, 2007)(Liao, Chu and Hsiao, 2012). Classification, association, clustering, & prediction are only a few of the purposes of DM. DM has emerged as one of the most critical and enticing areas of study for extracting useful information from massive amounts of data. Data mining has become popular especially in the field of healthcare since an appropriate technique for the discovery of unidentified and precious knowledge in the field of medical data is required. In healthcare, Data Mining has many advantages including detecting healthcare fraud, presenting patients with a reduced cost treatment approach, determining the risk factors, and identifying healthcare methods. This also assists health practitioners in designing successful health strategies, creating drug recommended systems, developing human health profiles, and so forth (Mitra, Pal and Mitra, 2002). DM has several benefits in the medical industry, such as identifying the separation of confidence in well-being safety, making restorative unions accessible to patients at low cost, also identifying causes of diseases, and compassion of therapeutic support systems (Tomar, 2013).

It is the goal of DM to uncover the hidden relationships and patterns in massive volumes of information contained in databases. It is possible to identify links between numerous attributes in a database's large amount of data and use this information to aid in many different decision-making processes. To discover these relationships, a technique called Association Rule Mining (ARM) was developed by (Agrawal and Srikant, 1994). Identification of these kinds of connections or patterns has become an exciting area of study due to the massive amounts of data being created every day. Market basket data analysis is a simple form of ARM, which uses correlations among products bought to evaluate the buying patterns of customers. Instead of relying on the data's implicit features, associations between items are determined by their presence together in the database (Koh and Rountree, 2009). ARM has been widely studied in the literature for mining these associations or relationships (Hipp, Güntzer and Nakhaeizadeh, 2000)(Agrawal et al., 1996).

The remaining paper is structured as follows: Section III talks about some applications of DM in health care. A comprehensive outline of the up-to-date methods for rare pattern mining has been discussed in Section III as a literature survey. Section IIV provided the details about the proposed methodology that describe the concept of minimum itemset support difference and also included the proposed MISDSSP-tree algorithm. Then experimental results and their evaluation has visualized in Section V with their descriptions. Finally, some research issues and future directions for this research on rare pattern mining have also been discussed in Section VI followed by the concluding remarks of this paper with future work in Section VII.

II. LITERATURE SURVEY

This section is intended to provide an analysis of all healthcare sectors based on day-to-day data. In medical research, it is the most urgent to examine diseases & the treatments of patients. As of late, manual recorded professionals' reports are turned into digital files, thus lowering costs during treatment and going to increase treatment performance. Data mining applications may also be applied to subsequent classifications of health security.

The field of Knowledge Discovery in Databases (KDD) is concerned with the advancement of methodologies & procedures for data utilization. DM is a critical phase in the KDD process. DM is a method for sifting through enormous volumes of data to find and extract patterns. Early detection systems, as well as other healthcare-related technologies, have been built using clinical and diagnostic data by both the DM and healthcare businesses, making use of the most trustworthy clinical and diagnostic data. In light of this development, we conducted a study of the numerous papers published in this area in terms of methodology, algorithms, and outcomes.

1) Rare Association Rule Mining (RARM)

We know that the primary goal of RARM is to identify unusual linkages or relationships between item sets in a transactional database.

To several application areas like the use of scam credit cards, network anomalous detection, network fault diagnosis, educational information, clinical issue, and so on, rare association rules can be advantageous. (Vanamala, Padma Sree and Durga Bhavani, 2014) applied the concept to evaluate the data stream to find interesting rules on rare associations. The method of defining associations with min sup but with higher confidence was RARM. This proposed rarity data stream ARM algorithm has been implemented that used a sliding window strategy, which displays the data in longitudinal bit sequence layout. The advantage of the presented algorithm is that the discovery of all rare associations requires single scanning only. For memory & time, the presented algorithm performs better.

Then, mining of utilities seeks to uncover items with the highest utilities, taking profit, number, cost or other user desires into account. In the retail industry, the usefulness of the item in the transactions should be weighed so products with a low frequency of sale will have the highest return. In various decision-making realms, rare item sets offer helpful information. (Niha and Dulhare, 2014) proposed a high utility rare item sets mining algorithm entitled UPRI (utility pattern rare item set), with the strategic approach for mine highest utilities. In a tree-based data system, entitled UPR-Tree, knowledge for the highest utilized item sets was retained. This work presented UPRI algorithms for generating rare item sets with high utilities. Such items are rare in a transactional dataset but can produce tremendous benefits for a corporation (Niha and Dulhare, 2014).

Per-SPAM means that frequent patterns are found in sequences of items. An MIS method for extracting rare items is Mis-SPAM. So, (Kaushal and Singh, 2016) presented 2 novels Per-SPAM & Mis-SPAM methods. Per-SPAM means that frequent patterns are found in sequences of items. An MIS method for extracting rare items is Mis-SPAM. The results obtained are the proposed New SPAM algorithm which implements the MIS & periodic method. Relying on their item supports, the novel algorithm assigns MIS values for frequent items & rare items. In comparison with frequency and runtime, the efficiency of the proposed algorithms and methods has been compared. the utility and encouraging success of the proposed framework.

MS Apriori uncovers the rare borderline itemset from weblogs as well as Dynamic Apriori uncovers the items that express a good association to the frequent items through ARM. (Kesarwani, Goel and Sardana, 2018) proposed the hybrid MSD-Apriori method to reveal the below, even so near to minimal support threshold borderline-rare items with the good association to frequent items. The hybrid method is combined with Dynamic Apriori by MS Apriori. Kosarak, a true dataset that provides impressive performance, analyzed the proposed hybrid method.

In this article, (Shrivastava and Dr. Varsha Jotwani, 2021) offered an efficient incremental approach for detecting rare patterns, namely, ISSP-tree algorithm that has used the minimum support difference for each itemset. The goal of this study is to create and construct a model that is efficient and productive for diagnosing adverse illness utilising rare patterns extraction.

2) Disease Identification using Rare Pattern/Rule Mining

In the medical field, the user will identify rare patterns or trends that help physicians' decisions on clinical treatment by reviewing clinical databases.

(Borah and Nath, 2018) applied the successful method to classify signs & risky factors in 3 adverse diseases: hepatitis, heart disease, and breast cancer with rare ARs. Once the database was changed, the consumer may move to a novel threshold value for desirable rare ARs. The new scheme throughout this study allows you to create new rare ARs in one database scanning from the latest clinical database without re-executing the whole mining process. It may manage transaction insertion and removal cases effectively and provide the user with versatility to create new rare ARs set when the thresholds are changed. Exploratory research reveals the importance of the method provided to the conventional method of the whole revised database constantly being undermined.

(Fournier-Viger et al., 2020) proposed to detect a new pattern type is called a rare correlated periodic pattern in multiples sequence. The issues were investigated and an effective algorithm called MRCPPS has been identified to discover such patterns effectively. Mining of Rare Correlated Periodic Pattern Common to Multiples Sequences (MRCPPS) A modern RCPPS-list layout has been used to keep the database from being searched again and again. Multiform experiments have been performed and found that the proposed MRCPPS algorithm was effective in identifying all the rare correlated frequent patterns common to multiples sequences, and in filtering numerous pattern which was not rare & related.

3) Adverse Disease based Association Rule Mining

Techniques from computational intelligence, like ARM, may be widely used in the investigation of unfavorable illnesses. Several of the primary disadvantages of association rule algorithms are that utilized algorithms have several parameters for someone who is not an expert in DM and that resulting rules are considerably too numerous, the majority of which are uninteresting also difficult to comprehend. In this chain, (Kabir, Ludwig and Abdullah, 2019) used the ARM technique for revealing the information in rule form using breast cancer data that may be beneficial in initiating preventive mechanisms. They found rules for

patients who have both breasts & also non-breast cancer so that both people realize but instead start comparing characteristics. The test outcomes reveal that produced or mined rules have a high confidence value. This research determined to estimate association b/w HEV & PTS. So, (Mendoza-Lopez et al., 2020) have identified two PTS cases linked to HEV which have been diagnosed in the same village within a short period. (Shrivastava and Jotwani, 2022) represents the comparison of various architecture research for classification of heart disease type using a different type of datasets. This study contributes to recognising the existing processes used in data mining to diagnose heart disease through classifications.

III. PROPOSED METHODOLOGY

In this section, we have described the proposed methodology. The problems have been identified in the existing work firstly then to overcome such problems by proposing a MISDSSP-tree algorithm. To begin, this work offers an efficient minimal item support difference pattern tree-based method for extracting a current set of unusual association rules from dynamically changing datasets. Two steps comprise the proposed approach: tree construction & pattern mining. In the first step, a compressed prefix-tree representation of the original database is produced in a single database scan. The second step generates substantial frequent & rare patterns from tree data structure without repeatedly referencing the original database.

A. Problem Statement

To extract rare items, an attempt has been done in literature work in which the percentages of each item are set equal to their percent. Although this method enhances efficiency over individual approaches, the "rare item issue" challenge remains. The rare item sets will be lost as the small for rare items gets closer to its support by adjusting the percent value to the high, and also the numbers of frequent items are exploded if min_sup for an item is adjusted by setting percent value to low.

In the existing work, we can see that the SSP-Tree algorithm was used which has several disadvantages:

- The SSP-Tree algorithm offers significant efficiency benefits with relatively minimal overhead tree restructuring for runtime & memory use. Since the method of a tree, restructuring requires some overhead computing due to partitions and node mergers.
- How to identify them effectively in vast dynamic datasets.
- How to Detect Interesting Rare Patterns.
- If the Min_sup value is set with a high % value then the rare items are ignored while the min_sup is near to support value for the rare items.
- If Min_sup value is set with a low % value then the no. of frequent item sets has been exploded.
- It also suffers from "rules missing" as well as "rules exploding" issues.
- In SSP-tree ordering has not been assigned to the items in the dataset.

Due to the above-addressed disadvantages, we opt to find the more efficient multiple minimal support-based dynamic RARM algorithm which can overcome the limitations of SSP-Tree.

B. Proposed Methodology

Here, we have introduced to use of the term support differences (SD) as a single scan pattern tree regarding the idea of minimum item support. The presented method is distinct from the SSP tree as it uses another way of assigning minimum support to each item. This proposed method encompasses frequent items with rare items only, while the proposed method extracts frequent items including frequent & rare items both. For this work, three real-time benchmarked datasets have been used to generate significant association rules & patterns of rare items.

1) Multiple Minimum Item Support Difference Approach

Each item is specified with the minimum support value named as minimal items support (MIS) in multiple minimum support value-based frequent patterns mining. Frequent items are such items whose support count is more than or equal to their corresponding MIS values. Items with less support count than their corresponding MIS values are not frequently called Rare [19].

This work has used the concept of SD for determining the min support for items for this method. The SD (support differences) means the appropriate divergence of an item from its occurrence (i.e. support count) to make an item set with such items a frequent itemset. For every item 'x_y,' the min item support formula ($MIS(x_y)$) seems to be as follows:

$$MIS(x_y) = Sup(x_y)\text{-}SD, \text{ when } (Sup(x_y)\text{-}SD) > LSup$$
$$= LSup, \text{ otherwise} \quad (1)$$

Here,

$Sup(x_y)$ = Item 'x_y' support value
LSup = User-defined least supports value.
SD= MIS value for all items ranges from $(-\infty, +\infty)$.

The SD value could be determined according to Eq.(2) for a given set of data.

$$SD = \lambda(1 - \alpha) \quad (2)$$

Here,
λ =Parameter that can be mean, mode, median, max support value for all item support
α = Parameter that have value (0 to 1).

In the determination of SD, λ & α play a significant part. Relevant statistical parameters can be determined to be optimal for the data set. The sampling techniques are used when the collected data size is immense. The user determines the α value. Here, we take the α value is 0.8 and λ is the maximum support count. SD contains $(0, \lambda)$ values.

After MIS values have been defined by Eq. (2) for each item, Eq. (3) is used to produce frequent itemset:

$$Sup(x_1, x_2,.., x_l,\ldots x_k) \geq Min(MIS(x_1),MIS(x_2),\ldots,MIS(x_k) \quad (3)$$

Here,

$Sup(x_1, x_2,\ldots,x_l,\ldots x_k)$= Itemset $\{x_1, x_2,\ldots,x_l,\ldots x_k\}$ support values

Eq. (2) conveniently retrieves a set of frequent items, including frequent items, rare items, and/or both rare items & frequent items. It should be remembered that the value of MIS depends on its support value for every item. When an item set comprises all frequent items, the least MIS of frequent items must be met. Likewise, if an item set includes frequent & rare items then the least MIS of the rare items shall be fulfilled in it. In this method, it can also be guaranteed that the MIS has been consistently differentiated from the support of an item regardless of its item support count values. The presented method is not likely to disperse or difference between support of the items so that it can be used for all types of data sets, such as data sets with widely differing item occurrences (i.e. supports) [20].

C. Proposed MISDSSP-Tree Algorithm

The proposed MISDSSP-Tree algorithm has been defined below:

Input: d=Diseases dataset (Cleveland heart, Breast cancer, Hepatitis), α=0.8, λ =max_count, LSup=15 min_rare=1, SD=(0, λ)

Output: Generate Frequent item sets & Rare item sets, No. of rules.

Procedure:

Assume CS_m labels the candidate set of m-itemset & FS_m labels the frequent set of m-itemset. Remember that items in a set of items are ordered from left to right in ascending order of MIS values. An algorithm to find the no. of rules and frequent & rare itemset is specified below.

Step 1. Produce candidate 1-itemset CS_1;
Step 2. Initially Calculate support count (Sup) to each Itemset in CS_1 and set the least minimum support count
Step 3. MIS=Compute-MIS(Sup, SD, LSup);
Step 4. FS_1={ x$\in$ CS_1, $Sup(x) \geq MIS(x)$};
Step 5. FS_1 = sort(FS_1,MIS);
Step 6. For m=2; $FS_{m-1}\neq\emptyset$;++m do
Step 7. CS_m=candidate-gen(FS_{m-1});
Step 8. For dataset d$\in$D do
Step 9. CS_d=subset(CS_d,d);
Step 10. For each candidate c$\in$$CS_d$ do
Step 11. {
Step 12. c.count++;
Step 13. }; // End for loop

Step 14. }; // End for loop

Step 15. $FS_m = c \in CS| \frac{c.count*100}{|d|} \geq Min(MIS(x)|\forall x \in c)$

Step 16. }; // End for loop

Step 17. Return=∪m FS_m;

Step 18. Calculate the minimum item support of each itemset by eq. (1)

Step 19. for x = 1; x ≤ | CS_1 |; ++x do

Step 20. {

Step 21. M(x) = Sup(x) − SD(n);

Step 22. If M(x) is less than LSup then MIS(x) equals to LSup otherwise MIS(x) equals M(x);

Step 23. }; // End for loop

Step 24. Return the value of MIS;

Step 25. In the same way, if a set of items includes frequent & rare items, the least MIS of rare items must have complied with it.

Step 26. Obtain frequent and rare patterns and many rules.

Step 27. Measure elapsed time and used memory.

Step 28. Exit

IV. Experimental Results & Evaluation

The experimental findings of this implementation work with data description are discussed in this section. This experimental simulation has done by the MATLAB tool. The simulation has been performed on three benchmarked datasets of adverse diseases and their description is also provided in this section. Performance of this proposed method is also measured in terms of taken time to execute & memory size to store it.

A. Dataset Description

Here, we have presented a short-term explanation of datasets. The dataset included here are the heart dataset, cancer dataset, and hepatitis dataset that are publicly online available at the UCI machine learning repository.

1) Heart dataset

The Cleveland Heart Disease dataset[1] is taken from the UCI ML repository. This dataset comprises 76 features but about 14 are only included in the cardiovascular disease diagnostic testing studies. Occurrence or absence of heart disease is shown by 303 instances & five class naming (0-4). The lack of class 0 and the cardiac condition of classes 1 to 4 was indicated. There are 164 sample cases in class 0, class1 having 55 samples, class2 having 36 samples, class3 having 35 samples, & class4 having 13 samples. The risk of heart disease is seen in classes 1 to 4. Table I refers to the features including values of the Cleveland heart dataset.

Table I. Information of features for Cleveland heart disease dataset

[1] *https://archive.ics.uci.edu/ml/datasets/heart+disease*

S. No	Features	Explanation	Data Type	Value
1.	Age	Patient Oldness	Numeric	29 to 77
2.	Sex	Gender	Binary	0= F, 1= M
3.	Chp	Types of chest pain	Nominal	1= Typical angina, 2= Atypical angina, 3= Nonanginal pain, 4= Asymptomatic
4.	Trestbps	Resting blood pressures	Numeric	94 to 200
5.	Ch	Cholesterol	Numeric	126 to 564
6.	Fbs	Fasting blood sugar greater than 120mg/dL	Binary	1= True, 0= False
7.	Restecg	Resting electro-cardiographics result	Nominal	0 = Normal, 1= Abnormal, 2= ST-T wave abnormality, 3= Left ventricular hypertrophy
8.	Thalach	Maximize obtained heart rate	Numeric	71 to 200
9.	Exang	Exercise involved angina	Binary	1 = Y, 0 = N
10.	Oldpeak	Exercise relatively to rest caused ST depression	Numeric	Continuous (from 0 to 6.20)
11.	Slope	Peak workout	Nominal	1= Upslope, 2= Flat, 3 = Downslope

		ST segment slope features		
12.	Ca	No.of fluoroscopy colored vessels	Nominal	0 to 3
13.	Thal	Types of defect	Nominal	3 = Normal, 6 = Fixed defect, 7 = Reversible defect
14.	Class	Healthy or heart disease existence	Binary	0= Safe, 1= Lower Risk, 2 = Medium Risk, 3 = Higher Risk, 4=Severe Risk

Note:- M: Male, F: Female, Y: Yes, N: No

2) Breast Cancer dataset

The Breast Cancer Wisconsin (Diagnostic) dataset[2] is also available at the UCI ML repository. It has 699 instances & 32 features. For diagnostic tests, about ten are included. Two classes of marks classify the types of breast cancer as benign and malignant. This has 458 sample cases in the benign & 241 sample cases in the malignant class. Wisconsin Breast Cancer data set feature information is available in table II.

Table II. Information of features for the Wisconsin breast cancer dataset

S. No.	Features	Range
I.	Clump thickness	one to ten
II.	Marginal adhesion	one to ten
III.	Cell size uniformity	one to ten
IV.	Bare nucleoli	one to ten
V.	Size of single epithelial cells	one to ten
VI.	Cell shape uniformity	one to ten
VII.	Normal nucleoli	one to ten
VIII.	Bland chromatin	one to ten

[2] *https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)*

IX.	Mitosis	one to ten
X.	Class	Benign or Malignant

3) Hepatitis dataset

The Hepatitis dataset[3] is also available at the UCI ML repository like two other datasets. There are 155 instances & 20 features in the dataset. The 'Died' & 'Alive' class labels signify whether or not such a patient who has hepatitis is going to be alive. There are 32 sample cases of 'Died' & 123 'alive.' The hepatitis data set has several null values. Table III displays the Hepatitis data set features information.

Table III. Information of features for the hepatitis dataset

S. No	Features	Explanation	Value
1.	Age	Patient Oldness	10 to 80
2.	Sex	Gender	M, F
3.	Steroid	Ingesting of steroid	N, Y
4.	Antivirals	Antiviral handling	N, Y
5.	Fatigue	Sign	N, Y
6.	Malaise	Sign	N, Y
7.	Anorexia	Sign	N, Y
8.	Liver firms	Solid liver	N, Y
9.	Liver big	Widen liver	N, Y
10.	Spider	Spider nevus	N, Y
11.	Spleen palpable	Distended spleen	N, Y
12.	Ascites	Liquid b/w stomach & tissues	N, Y
13.	Varices	Widen vein	N, Y
14.	Bilirubin	Quantity of the bilirubin	0.390 to 4.000
15.	Sgot	Liver enzyme	13 to 500
16.	Alk phosphate	Quantity of an alkaline phosphate	33 to 250
17.	Albumin	Quantity of an albumin	2.10 to 6.00
18.	Protime	Protein for liver	10 to 90
19.	Histology	Histology of liver	N, Y

[3] *https://archive.ics.uci.edu/ml/datasets/hepatitis*

20.	Class	The patient is dead or alive	Died, Alive

B. Results and Analysis

In terms of no. of items sets & produced rules, the efficacy of the proposed MISDSSP-Tree method regarding previous methods could be assessed. MISDSSP-Tree method is also able to extract a significant number of frequent & rare item sets. Table IV presented produced no. of item sets by the proposed MISDSSP-Tree method and existing SSP-tree method from the Cleveland Heart Disease datasets respectively.

Table IV. No. of Item sets produced for heart disease dataset

Approach	Generated Frequent Item sets	Generated Rare Item sets
SSP-Tree (Existing)	486	976
MISDSSP-Tree (Propose)	413	973

Except for existing SSP-tree configurations, the MISDSSP-Tree configuration preserves all database items, providing the user with the flexibility to create any number of desirable patterns.

Table V. Item sets Generated from breast cancer dataset

Algorithm	#Frequent Item sets	#Rare Item sets
SSP-Tree (Existing)	298	2229
MISDSSP-Tree (Propose)	2066	2138

MISDSSP-Tree method has also started to achieve a full set of rare ARs & item sets on Wisconsin Breast Cancer datasets as well represented in Table VI.

Table VI. Produced rare rules from the breast cancer dataset using proposed MISDSSP-Tree

Rules	Antecedent	Consequence	Supp (%)
1	Bare Nucleoli=1	Size of Single Epithelial Cells =9	0.0029
2	Size of Single Epithelial Cells =3	Size of Single Epithelial Cells =9	0.0029

3	Marginal Adhesion=7	Size of Single Epithelial Cells =9	0.0029
4	Marginal Adhesion=7 AND Size of Single Epithelial Cells=3 AND Bare Nucleoli=1	Size of Single Epithelial Cells =9	0.0014
5	Bare Nucleoli=1	Mitoses=6	0.0029
6	Bare Nucleoli=1	Mitoses=6	0.0029
7	Size of Single Epithelial Cells =3	Mitoses=6	0.0029
8	Size of Single Epithelial Cells=3 AND Bare Nuclei=1 AND Bare Nucleoli=10	Mitoses=6	0.0014
9	Mitoses=1	Marginal Adhesion=9	0.0043
10	Bare Nucleoli=10	Marginal Adhesion=9	0.0086

The rare ARs produced by the proposed MISDSSP-Tree are shown in Table VIII. Rule1 holding the rare item "Bare Nucleoli=1", specifies that once the size of the single epithelial cells is assigned 9 as a value at support 0.0029 through the physicians, this indicates a normal tumor.

Table VII. No. of Item sets produced for Hepatitis dataset

Algorithm	#Frequent Item sets	#Rare Item sets
SSP-Tree (Existing)	436	918
MISDSSP-Tree (Propose)	213	836

From Table VIII, it is obvious that the MISDSSP-Tree method can output a significant no. of item sets & rules for the Hepatitis dataset, as compared to another existing SSP-tree algorithm. Thus, the MISDSSP-Tree algorithm provides significant patterns that are required to detect Hepatitis disease based on attributes in terms of no. of item sets & rare ARs.

Table VIII. Produced rare rules from the Hepatitis dataset using proposed MISDSSP-Tree

Rules	Antecedent	Consequence	Supp (%)

1	Histology=Yes	Sgot=[489.6-648]	0.0129
2	Albumin=[2.1-3.175]	Sgot=[489.6-648]	0.0129
3	Anorexia=No	Sgot=[489.6-648]	0.0129
4	Malaise=No	Sgot=[489.6-648]	0.0129
5	Fatigue=No	Sgot=[489.6-648]	0.0129
6	Antivirals=Yes	Sgot=[489.6-648]	0.0129
7	Steroid=No	Sgot=[489.6-648]	0.0129
8	Sex=Male	Sgot=[489.6-648]	0.0129
9	Sex=Male AND Steroid=No AND Antivirals=Yes AND Fatigue=No AND Malaise=No AND Anorexia=No AND Albumin=[2.1-3.175] AND Histology=Yes	Sgot=[489.6-648]	0.0065
10	Class=Live	Age=[0-20]	0.0194

The rare ARs produced by a MISDSSP -Tree is given in Table IX. Rule1 taking the first rare item "Histology=Yes", indicates that people with liver histology have had chances of recovery to Liver enzyme at support 0.0129 from hepatitis. According to rule 2 taking the rare item, "Albumin=[2.1-3.175]" people having Sgot at support 0.0129 have the lowest Liver enzyme for hepatitis. Similarly, for all rules, this works. Here in table X, we took 10 rules from the complete set of rules.

Table IX. Elapsed time (in seconds) taken for Heart, Breast cancer, and Hepatitis dataset

Algorithm	Heart dataset	Breast cancer dataset	Hepatitis dataset
SSP-Tree	58.356289	1367.991366	11.801659
ISSP-Tree	52.854660	1353.420129	11.225005
MISDSSP-Tree	50.238063	1381.054007	11.667285

Table IX represents the elapsed time for all SSP-Tree, ISSP-Tree, and MISDSSP-Tree methods on the Heart, Breast cancer, and Hepatitis dataset. From table XI, we can see that the ISSP-Tree takes time 52.85 seconds, 1353.42 seconds, and 11.23 seconds for the Heart, Breast cancer, and Hepatitis dataset, respectively, that are less than SSP-tree while MISDSSP-Tree takes time 50.24 seconds, 1381.05 seconds, and 11.67 seconds for the Heart, Breast cancer, and Hepatitis dataset, respectively, which are also less than SSP-tree but it is high than ISSP-Tree method for all datasets except Heart disease dataset.

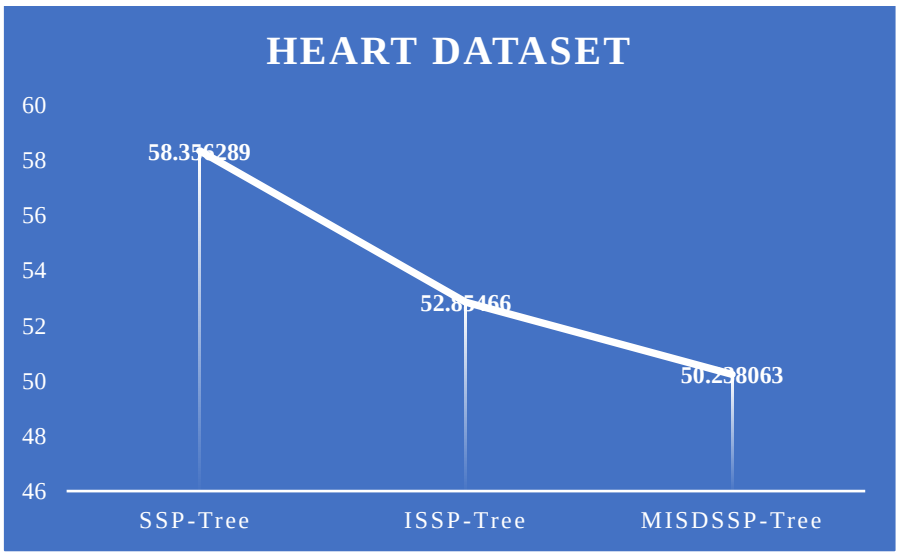

Fig. 1. Elapsed time (in seconds) taken for Heart dataset

Fig. 1 shows the runtime consumed by frequent patterns and rare patterns for different three datasets. Heart data set insertion demonstrates the individual impact for frequent & rare pattern mining in figure 1. Execution time invested by SSP-tree proposed ISSP-tree algorithm for generating frequent & rare patterns in several support count values is represented in Figure 1.

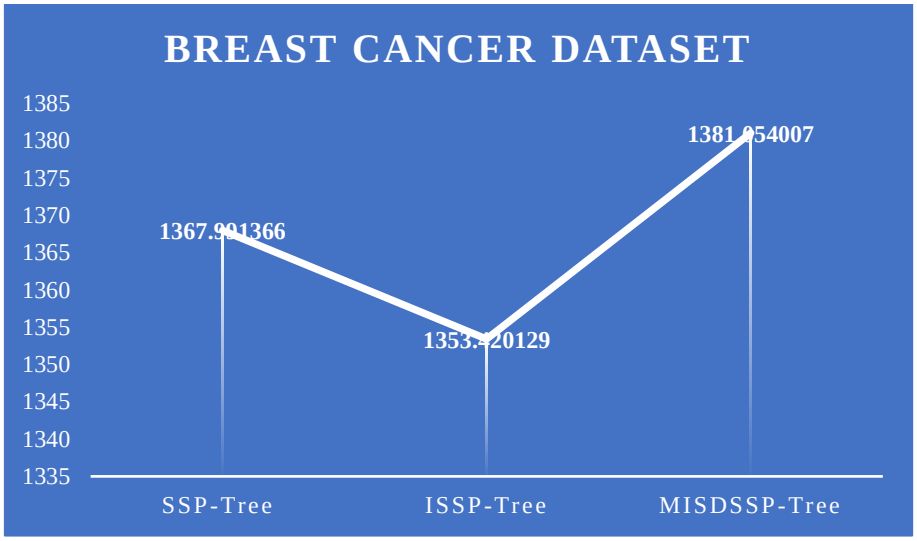

Fig. 3. Elapsed time (in seconds) taken for Breast cancer dataset

Fig. 3 shows the runtime consumed by frequent patterns and rare patterns for different three datasets. Breast cancer data set insertion demonstrates the individual impact for frequent & rare pattern mining in fig. 3. The time is calculated in seconds. The proposed MISDSSP-tree takes high elapsed time in comparison to the existing SSP-tree while ISSP-tree takes very minimal time in comparison to both approaches to the Breast cancer dataset.

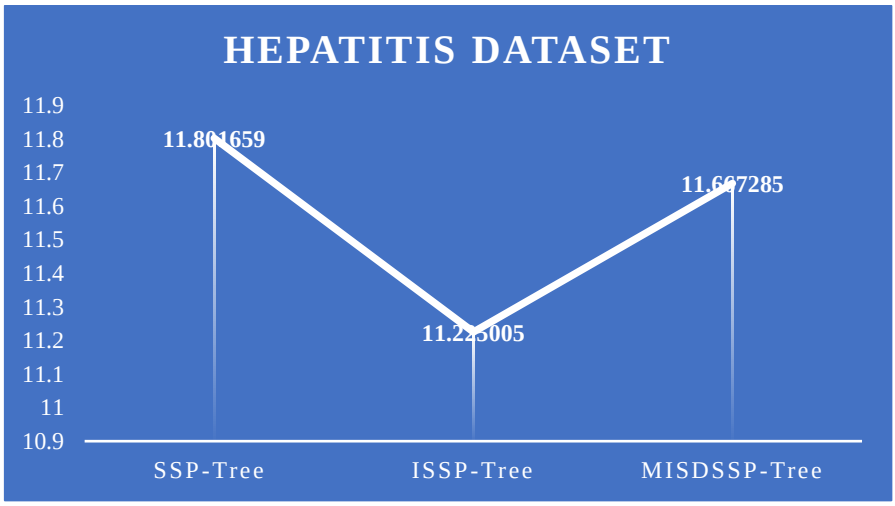

Fig. 3. Elapsed time (in seconds) taken for Hepatitis dataset

Similarly, Fig. 3 shows the runtime consumed by frequent patterns and rare patterns for different three datasets. Hepatitis data set insertion demonstrates the individual impact for frequent & rare pattern mining in fig. 3. The time is calculated in seconds represents for all approaches. The proposed MISDSSP-tree takes high elapsed time in comparison to the ISSP-tree but takes less than the existing SSP-tree to Hepatitis dataset.

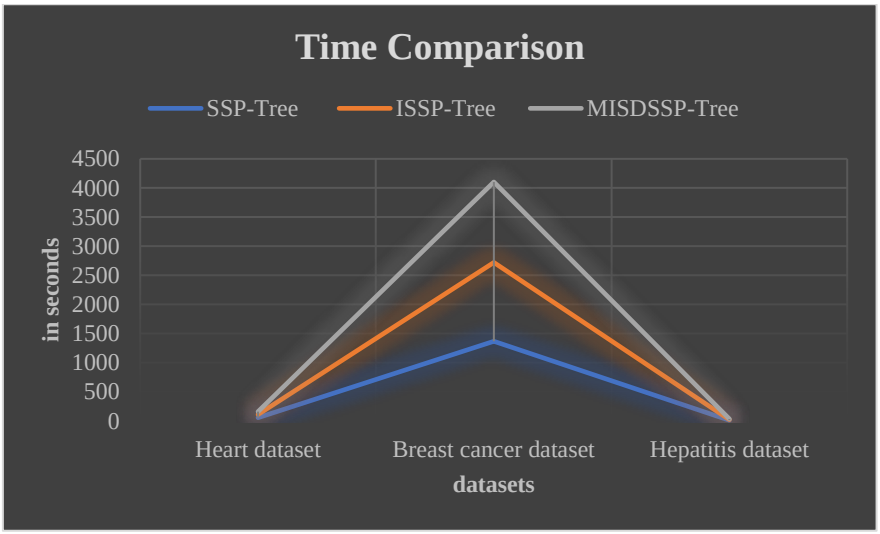

Fig. 4. Elapsed time comparison line graph

Fig. 4 depicted the elapsed time taken by rare & frequent patterns for all three datasets. Every data set insertion demonstrates the individual impact for frequent & rare pattern mining in fig. 4. The proposed MISDSSP-tree minimized the elapsed time in comparison of the existing SSP-tree and ISSP-Tree to all three datasets namely, the Heart, Breast cancer, and Hepatitis dataset. The x-axis represents datasets & the y-axis shows time in seconds.

Table X. Memory Used (in KB) taken for Heart, Breast cancer, and Hepatitis dataset

Algorithm	**Heart dataset**	**Breast cancer dataset**	**Hepatitis dataset**
SSP-Tree	3801.243164	5228.993164	2370.102539
ISSP-Tree	3559.692383	4533.442383	1977.977539
MISDSSP-Tree	3466.524414	4307.989258	1980.526367

Table X represents the memory consumption for all SSP-Tree, ISSP-Tree, and MISDSSP-Tree methods on the Heart, Breast cancer, and Hepatitis dataset. From table XII, we can see that the ISSP-Tree used memory 3559.69 KB, 4533.44 KB, and 1977.98 KB for Heart, Breast cancer, and Hepatitis dataset, respectively, that are much less than SSP-tree while MISDSSP-Tree used memory 3466.52KB, 4307.99KB, and 1980.53KB for Heart, Breast cancer, and Hepatitis dataset, respectively, that are also much less than the existing SSP-tree and ISSP-Tree as well for all datasets except Hepatitis dataset.

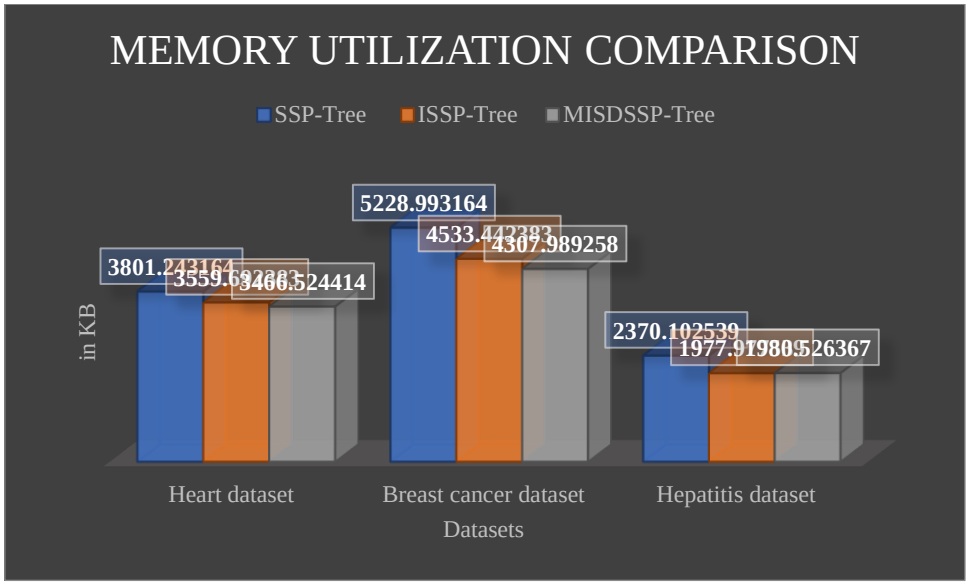

Fig. 5. Utilized memory comparison bar graph

The memory usage of the existing SSP-tree, ISSP-tree, and proposed MISDSSP-tree for generating a frequent and rare pattern on Heart, Breast cancer, and Hepatitis dataset has illustrated in fig. 5. The memory size is taken in Kilo Byte (KB). Fig. 5 indicates memory usage in various upgrade sizes & support thresholds for frequent & rare pattern generation. In terms of efficient Memory usage, the proposed MISDSSP-tree algorithm for all data sets has outperformed. In terms of efficiency, both proposed ISSP-tree and MISDSSP-Tree for all data sets still outperformed. But MISDSSP-Tree is much more memory efficient in comparison to the existing SSP-tree and proposed ISSP-tree. The breast cancer dataset required more memory usage compared to the other two datasets, while the hepatitis dataset consumed less memory. However, the hepatitis dataset required less memory in contrast to the other two datasets but MISDSSP-Tree consumed quite more memory compared to ISSP-tree.

V. FUTURE DIRECTIONS

Handling incremental datasets: In the case of dynamically updated datasets, a database may produce new rules that invalidate previous ones when they are modified (Nath, Bhattacharyya and Ghosh, 2013). As a result, it is difficult to keep up with the rules of the association. An alternative strategy for dealing with the issue might be to run a mining algorithm again on updated data. However, even though this strategy is straightforward, it is ineffective because it does not take into account the number of items in a big item set that are now supported. As a result, it is necessary to use algorithms that act just on the newly acquired data, rather than beginning from scratch. For incremental frequent pattern mining, there are several methods available. While there has been some progress in incremental rare pattern mining, considerable work has to be done. Scalable techniques: Frequent itemset mining relies heavily on the amount of physical memory available. Real-world databases are becoming larger and larger, and this has made primary memory size a major concern for present solutions (Adnan and Alhajj, 2009).Many heuristic-based ARM strategies have been published, but only one has been implemented in the RARM field (Luna, Romero and Ventura, 2014).Mining rare patterns from sequential

databases: Numerous difficulties arise in the field of sequential pattern mining (Manohar, Dinesh and Sowmya, 2015). Attempts were made by (Zhu et al., 2016) to mine document streams for rare sequential patterns. Sequential topic patterns (STPs) is an algorithm that attempts to detect aberrant behavior on the internet using document streams. Despite its potential, the mining of rare sequential patterns has yet to be fully investigated.

VI. CONCLUSION

Across the globe, the rising death toll from preventable diseases has become a serious issue. Using computational intelligence approaches such as RARM, the rare connections between distinct patient traits and diseases may be studied to improve decision-making & medical diagnosis. Because of their pervasiveness, CVD, hepatitis, & breast cancer are three of the most important causes of death that this research seeks to investigate. This research presents an effective method for creating a list of relevant rare association rules from 3 clinical datasets: Hepatitis, Wisconsin Breast Cancer, & Cleveland Heart Disease datasets are all available online. The medical specialists have been supplied with a complete and exhaustive study of the extracted unusual association rules to aid in the identification of these terrible diseases. Existing RARM algorithms, on the other hand, only work on static databases and repeat the whole mining process if the original database is altered. We have constructed a support difference-based technique called MISDSSP-Tree in this article. The proposed MISDSSP-Tree method is capable of handling all aspects of a dynamic database, including transaction insertion, deletion, & threshold updates. To build the MISDSSP-Tree structure, it does a single scan of the database, which is dynamically restructured each time a new transaction is added or removed. The approach avoids repeatedly scanning the database and reconstructing the tree structure, thereby successfully extracting the rare association rules from dynamic datasets. Using three commonly used real-life clinical datasets, the proposed technique is contrasted to other pattern mining algorithms for efficiency. The proposed approach was compared to the existing SSP-tree and ISSP-tree algorithms in terms of runtime and memory utilization. A comparative study of results demonstrates the superiority of the proposed technique over current frequent & rare pattern mining techniques and its ability to generate substantial uncommon association rules relating to medical diagnosis. The MISDSSP-Tree technique offers a significant performance boost in terms of memory usage & execution time with relatively small tree restructuring overhead.

References

Adnan, M. and Alhajj, R. (2009) ‘DRFP-tree: Disk-resident frequent pattern tree’, Applied Intelligence. doi: 10.1007/s10489-007-0099-2.

Agrawal, R. et al. (1996) ‘Fast discovery of association rules’, Advances in knowledge discovery and data mining.

Agrawal, R. and Srikant, R. (1994) ‘Fast Algorithms for Mining Association Rules’, in Proc. of 20th International Conference on Very Large Data Bases, {VLDB’94}.

Borah, A. and Nath, B. (2018) ‘Identifying risk factors for adverse diseases using dynamic rare association rule mining’, Expert Systems with Applications. doi: 10.1016/j.eswa.2018.07.010.

Fournier-Viger, P. et al. (2020) ‘Discovering rare correlated periodic patterns in multiple sequences’, Data

and Knowledge Engineering. doi: 10.1016/j.datak.2019.101733.

Hipp, J., Güntzer, U. and Nakhaeizadeh, G. (2000) 'Algorithms for association rule mining — a general survey and comparison', ACM SIGKDD Explorations Newsletter. doi: 10.1145/360402.360421.

Kabir, M. F., Ludwig, S. A. and Abdullah, A. S. (2019) 'Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining', in Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018. doi: 10.1109/BigData.2018.8622028.

Kaushal, C. and Singh, H. (2016) 'New algorithm for finding frequent and rare itemsets', in 12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015. doi: 10.1109/INDICON.2015.7443264.

Kesarwani, S., Goel, A. and Sardana, N. (2018) 'MSD-Apriori: Discovering borderline-rare items using association mining', in 2017 10th International Conference on Contemporary Computing, IC3 2017. doi: 10.1109/IC3.2017.8284319.

Koh, Y. S. and Rountree, N. (2009) Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection. doi: 10.4018/978-1-60566-754-6.

Kumar, N., Jain, S. and Chauhan, K. (2019) 'Knowledge Discovery from Data Mining Techniques', International Journal of Engineering Research & Technology, 07(12), pp. 1–3.

Liao, S., Chu, P. and Hsiao, P.-Y. (2012) 'Data mining techniques and applications - A decade review from 2000 to 2011', Expert Syst. Appl., 39, pp. 11303–11311.

Luna, J. M., Romero, J. R. and Ventura, S. (2014) 'On the adaptability of G3PARM to the extraction of rare association rules', Knowledge and Information Systems. doi: 10.1007/s10115-012-0591-9.

Manohar, M., Dinesh, R. and Sowmya, M. S. (2015) 'Sequential Pattern Mining (SPM) for user-inputted data sets: an empirical framework using bitwise operations', International Journal of Knowledge Engineering and Data Mining. doi: 10.1504/ijkedm.2015.074083.

Mendoza-Lopez, C. et al. (2020) 'Parsonage-Turner syndrome associated with hepatitis E infection in immunocompetent patients', Virus Research. doi: 10.1016/j.virusres.2020.198165.

Mitra, S., Pal, S. K. and Mitra, P. (2002) 'Data mining in soft computing framework: a survey', IEEE Transactions on Neural Networks, 13(1), pp. 3–14. doi: 10.1109/72.977258.

Nath, B., Bhattacharyya, D. K. and Ghosh, A. (2013) 'Incremental association rule mining: A survey', Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. doi: 10.1002/widm.1086.

Niha, S. A. R. and Dulhare, U. N. (2014) 'Extraction of high utility rare itemsets from transactional databases', in International Conference on Computing and Communication Technologies, ICCCT 2014. doi: 10.1109/ICCCT2.2014.7066754.

Shrivastava, K. and and Dr. Varsha Jotwani (2021) 'ISSP-Tree: Minimum Item Support Based Improved Single Scan Pattern Tree for Generating Dynamic Frequent and Rare Patterns', Journal of education, XXIV(2), pp. 122–131.

Shrivastava, K. and Jotwani, V. (2022) 'A Comparative Analysis of Various Data Mining Techniques to Predict Heart Disease', in Jeena Jacob, I. et al. (eds) Expert Clouds and Applications. Singapore: Springer Singapore, pp. 283–296.

Sumathi, S. and Sivanandam, S. N. (2007) 'Active data mining', Studies in Computational Intelligence. doi: 10.1007/978-3-540-34351-6_13.

Tomar, D. (2013) 'A survey on Data Mining approaches for Healthcare', International Journal of Bio - Science and Bio - Technology, 5, pp. 241–266. doi: 10.14257/ijbsbt.2013.5.5.25.

Vanamala, S., Padma Sree, L. and Durga Bhavani, S. (2014) 'Rare association rule mining for data stream', in International Conference on Computing and Communication Technologies, ICCCT 2014. doi: 10.1109/ICCCT2.2014.7066696.

Zhu, J. et al. (2016) 'Mining User-Aware Rare Sequential Topic Patterns in Document Streams', IEEE Transactions on Knowledge and Data Engineering. doi: 10.1109/TKDE.2016.2541149.