A Model to discover Rare patterns using ML techniques

Deeksha Ramakrishna
Otto Von Guericke University
Magdeburg, Germany
deeksha.ramakrishna@st.ovgu.de

Stanley C. Umeh
Otto Von Guericke University
Magdeburg, Germany
stanley.umeh@st.ovgu.de

Surabhi Katti
Otto Von Guericke University
Magdeburg, Germany
surabhi.katti@st.ovgu.de

Seles Selvan
Otto Von Guericke University
Magdeburg, Germany
seles.selvan@st.ovgu.de

Priyamvada Bhardwaj
Otto Von Guericke University
Magdeburg, Germany
priyamvada.bhardwaj@st.ovgu.de

***Abstract*—Discovering Rare patterns from data is proving to be of tremendous importance in current data-driven world. These uncommon, hidden patterns are helpful in solving a range of crucial real world problems such as online security, malicious behaviour etc. While a few rare itemset generation methods adopt outlier detection approach using unsupervised clustering techniques, others employ variations of support threshold values. Current approaches to finding rare patterns focus on better performance but lack interpretability. RP-tree based RP-growth algorithm surpasses existing rare itemset generation algorithms both in terms of quality of rules and performance. This paper compares the rare rules generated by RP-Growth algorithm to one from the density based clustering approach. We shows that by utilizing outlier detection techniques in combination with association rule mining, the rare rules discovered are more meaningful and easily interpretable. We also show that rare rules generated using clustering approaches are more helpful in data imbalance problems. These rare rules extract reliable hidden patterns in minority classes, unlike imprecise rules generated by RP-Growth.**

***Index Terms*—Rare Pattern Mining, Association Rules, Imbalanced data**

I. Introduction

Frequent Itemset Mining (FIM) is a core concept in data mining which helps discover meaningful and informative knowledge from the data. FIM aims at discovering frequent items that co-occur in a dataset. The initial FIM studies were proposed to examine the customer behavior in real life problems, such as Market Basket Analysis. Using association rules [7] relationship between frequently occurring itemsets is expressed adequately. For instance, the problem of predicting a fatal disease for a patient based on available symptoms/conditions can be solved using FIM. FIM can discover frequently co-occurring symptoms/conditions resulting in the disease. However, frequent patterns might be less interesting than hidden rare patterns.

Considering the disease prediction example given above, there may be two notable problems: a) Annotated data collection for rare diseases is not only difficult, but can be prohibitively expensive as well b) Predicting that the user does not have the disease when they in reality have it is catastrophic. This under-representation of the minority class leads to imbalanced data, which is also true for problems like fraud detection, cyber attacks etc. Another reason for lack of data for such scenarios is that rare events are by definition, rare, which makes it extremely difficult to answer questions like ”what common relationship of items exists for this event?”. This in turns leads to the problem of data generalisation, any patterns derived from such an imbalanced dataset may have no intrinsic value for real-life scenarios. These shortcoming of finding only frequent itemsets demands the mining of rare itemsets. Finding rare rules not only helps with minority data, but is also helpful in refining the existing captured relationships where data is in plenty.

Finding rare patterns is beneficial in solving and refining existing solutions to real-life problems such as online payment fraud, equipment failure [26] [27] etc. Current approaches to finding rare patterns still have a number of open issues [2], user-defined support threshold is one of them. Since support threshold of an item sets the ratio of its occurrence in data, rare itemsets often do not meet the high support threshold even when data is in abundance. Hence, the minimum support threshold needs to be kept very low in order to capture rare patterns. Low support threshold leads to a number of problems such 1) Data generalization 2) Explosive growth in number of rules generated, and performance implications associated with it 3) Difficulty in determining if the rare rules generated are

genuine or random noise.

State of art methods employ different approaches to solve the problem of generating rare itemsets. Few methods use an unsupervised machine learning method such as clustering [3]. Some of the methods use additional user-defined support threshold, e.g. RP-growth [11] uses additional rare minimum support parameter. Other methods use tree-based approaches to combat performance issues, e.g. RP-Tree. All the approaches however focus on generating only rare rules. We aim to generate both rare and frequent rules for better interpretability. Clear distinction of frequent and rare rules can help get better insights from the data. Comparability of both type of rules makes rare rules more reliable, meaningful and useful.

The contributions of this paper are as follows:

- Our first contribution in this paper is to show that clustering based approach usually gives better results compared to multiple support based approach RP-Growth. Generating segregated clusters of frequent and rare rules not only helps users with domain knowledge to further validate and compare the rules, it also makes the results more intuitively understandable for those without any domain knowledge.
- The second contribution is that we extend and enhance existing clustering based approach for better performance and quality of generated rules.

The remainder of this report is summarized as follows: Section II provides the background information about the topic. Background information is further divided into two parts. Part A covers the basics of clustering and association rules. Part B covers similar work done on the topic. In Part III, Methodology, the clustering method adopted in the paper is explained. Methodology is divided into further sections about generating rules and clustering techniques. Evaluation strategy, performance criteria, and variations are discussed in Part IV. We discuss the results and observations in Part V. At the end of the paper, the conclusion is discussed as Part VI.

II. Background

A. Preliminaries

To define the rare and interesting rare itemsets formally we borrowed the definition from the papers [11] and [20] as follows:

Definition 1.1" (Rare Item set). Given a user-specified minSup threshold $minsup \in [0, 1]$, X is called a rare itemset or rare pattern in D if sup(X, D) $\leq minsup$."

Definition 1.2 "Interesting rare item sets: An itemset X whose support satisfies the following conditions is called the interesting rare itemset: $Sup(X) < maxSup \wedge Sup(X) \geq minSup$."

D is the total number of transactions in a dataset. Support (Sup) of an item set is counted as how many times the item set has occurred out of total number of transactions. Maximum Suppport (*maxSup*) and Minimum Support (*minSup*) are the upper and lower limit of support threshold respectively.

Clustering is used in unsupervised machine learning to club together similar points. There can be many different clusters containing points, each cluster with points that have similar characteristics. Points that do not belong to any of the clusters are called noise points/outliers. Similarity of data points can be computed based on various distance/similarity measures such as cosine similarity, Euclidean distance, etc. In density-based clustering method, points that are closer to each other are put into same cluster. Minimum number of points required to form a cluster can be defined with parameter *minPts*. Radius of a point to define its neighbourhood can be set using *eps* parameter. All the points within the neighbourhood of a point belong to the same cluster. A core point is a point when there are at least *minPts* within its radius *eps*, including itself. A neighborhood point differs from a core point in that it does not have *minPts* within its radius, but falls within *eps* of a core point. The rest of the points that are neither core nor neighbor points are noise points, and fall in a less dense region.

B. Related Work

Rare itemsets mining has great potential in solving real-life problems. Recently, a variety of research work is gaining speed in recovering rare itemsets. There are already a plethora of methods presented such as [27], [26], [25], [23], [19], [18], [8], [9] [15], [13] etc. We present some of the notable work in three high-level categories sequentially in upcoming segments: Support threshold based, Tree based, and Clustering based. In the first section we introduce support-based methods which are focused on mining rare itemsets, specifically by refining the support threshold requirement. Within support based approach a subsection implements variations of the most basic and successful Apriori algorithm. These Apriori based methods includes algorithms such as MSApriori [10] which allows each itemset in data to have its own minimum support, and Apriori-Inverse [25] captures all itemsets below a support threshold. Rest of the methods focus on tweaking the support requirements exclusively. In [19] authors assign weights to each itemset to estimate its minsup, LPMiner [15] algorithm uses constraint-based approach, ARIMA [13] algorithm is a mixture of three different algorithms, Rarity algorithm [23] performs the top-down search, since rare itemset tends to be on top in level-wise approaches. All support threshold approaches have drawbacks of generating redundant rules and suffer from performance issues.

On the other hand, a section of methods employs tree data structure to combat the performance issues. Tree based approaches including RP-Tree [18], Inverse FP-Tree, RPP algorithms [11] etc. are better performing in run time. The RPP algorithm uses the novel data structure RS list of all interesting rare item sets, and omits useless candidate item sets. However, if the data is not static but coming in streams, the rule generation process can be delayed, and the rules generated can be inconsistent. Additionally, the problem of determining the validity of the generated rules still remains, as many of them may just be random noise rather than a rare rule. Another interesting state of the art technique which uses the FP-tree method is Negative Itemset Tree [16], which basically takes the negation of the original item set. However the authors

warns of the limitation that it is unsuccessful on sparse data set.

There is another point of view to look at the rare rules generation which mirror it to Outlier detection used in unsupervised clustering methods [6] [5]. For example, [1] and [3] proposed clustering techniques to find rare rules. Other ideas in this direction include implantation of this approach in road accident detection [14] and in health care for data summarizing [9]. Interestingly, paper [5] also extended the use of adversary training which tries to find the key defining features of the object in order to find outliers. The state of the art work [4] uses clustering techniques called DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to segregate the frequent rule cluster from the rare rule cluster. Although this paper generates the informative rare rules, it uses Apriori algorithm to do so. Apriori algorithm has the drawback of generating huge number of rules with reduced run time performance because of the candidate generation approach it takes to generate the rules. The DBSCAN method results also vary with the choice of hyper parameters.

III. Methodology

Inspired by the application of clustering methods and negative item tree approaches, we propose methods to fill the gap in current techniques. [20] [16] [23] [4]. The overview of the existing approach is as below.

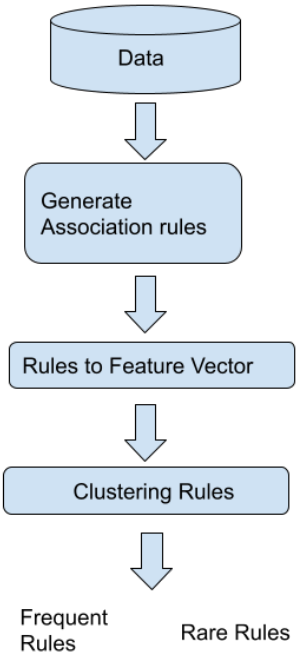

Fig. 1. Approach of generating frequent and rare rules

The method first generates the association rules capturing all itemsets, converts these rules into vectors based on co-relation of itemsets, and then applies clustering techniques to cluster frequent and rare rules. We divide the approach into 2 sections:

A. Generating Rules

Authors of paper [4] generate the association rules using Apriori algorithm. To be able to capture the rare itemset, the min support threshold has to be kept very low which results in slow run time performance.

In this paper, we consider the following algorithms in order to improve run time performance of generating association rule:

1) FP-Growth algorithm
2) ECLAT(Equivalence Class Clustering and bottom-up Lattice Traversal) algorithm

The FP-growth algorithm uses FP-tree to hold frequent itemset information, unlike candidate generation in Apriori algorithm. FP-growth algorithm scans the database at most twice while Apriori algorithm needs to pass over data multiple times. ECLAT algorithm uses a depth-first search like approach to generate the frequent itemsets. Both algorithms are faster, scalable, and more efficient than the Apriori algorithm [22], because there is no candidate generation step in FP-Growth algorithm. Both FP-Growth and ECLAT algorithms use tree-based structure to support compact information storage. ECLAT algorithm works in a vertical manner and does not require multiple passes over the data. Candidate generation and multiple passes over data used by Apriori algorithm make it costly in terms of time and memory.

B. Clustering Method

Authors in [4] chose the density-based DBSCAN (Density-based spatial clustering of applications with noise) algorithm to cluster the rules. DBSCAN algorithm automatically discovers clusters of rules. Rules with higher co-relation to other rules based on its itemsets stays in the same cluster. Few rules with weak co-relation and same itemsets are marked as noise by DBSCAN. Weaker co-relation results in more distance between the rules. These rules do not belong to other clusters as well because of different itemsets. In spatial terms, rules with similar itemsets and high co-relation forms dense regions while noise rules falls in a low density region. In DBSCAN algorithm hyper parameter *eps* determines how much a cluster would expand. A larger value of *eps* can include dissimilar rules into the cluster and a smaller value might not find a relevant cluster. Hence, choice of *eps* directly affects the results of the clustering process. Authors employs two additional parameters delta1 and delta2 for contradiction check between clusters, increasing the number of hyper parameters to tune. *Delta1* and *Delta2* parameters puts a constraint on *eps* value so that no rule with a direct contradiction is part of the cluster. The contradiction check function is described at the end of the section.

Another limitation of the DBSCAN comes to the picture when data points have different densities in a cluster. Since the value of *eps* is static, only the points satisfying *eps* condition become part of the cluster. Rules which are farther than *eps* but still belong to same cluster because of same itemsets cannot be found. Hence, the nested cluster structures cannot be captured.

We propose another density-based algorithm *Optics* to overcome these limitations. Optics algorithm sorts the data points based on the reachability distance of a point p to all other points. Reachability distance is the largest value between the distance between two points and the point's core distance. The core distance of a point is the largest distance between its n neighbors, where n is *minPts*. This enables the data points which are within the reachability distance of each other to stay together. The concept of reachability distance reduces the necessity of providing *eps* value since it is automatically calculated based on *minPts*. This helps in dealing with different densities within a cluster and helps discover nested cluster structures. Parameter *eps* can be set in Optics to reduce runtime; however, only required hyperparamter remains *minPts*, making it more straightforward.

Clustering needs vectors as input to compute the distance between points. Association Rules are converted to feature vectors. Number of unique features/itemsets defines the length of a feature vector. A co-relation matrix between itemsets using Spearman rank-order correlation is computed. The co-relation value of each item in a rule w.r.t to all other items defines feature vector of the rule. Rules containing more than one item are converted to feature vectors using the same technique except the co-relation of all items in the rule is compared to all unique items and the value of maximally co-related item is used.

Paper [20] introduced an interesting approach regarding categorising noise points discovered from clustering algorithms. Authors called it a- "Contradiction check" and used it to refine the clustering process. Refinement based on contradiction check added two additional hyper parameters to tune in clustering process. We propose a process to use the contradiction check function to validate the noise rules after clustering instead of putting an extra constraint on clustering process. Since Optics method automatically discovers the optimal *eps* value, contraction check function with additional parameters proves to be redundant. Instead contradiction check can be used to access the quality of rules and comparability. As per the check, only the fact that clustering process marks a rule as a noise point is not sufficient to declare it as a rare rule. The rule could be marked as noise because of ineffective parameters. The approach validates each noise rule against all other existing rules for following steps defined. A marked noise rule that satisfy all of the following conditions is considered a rare rule.

If there exists two rules X→Y and X' →Y' where

1) Y != Y'
2) Cosine similarity of X and X' is high
3) Confidence of both the rule is high
4) Rule X→Y is a noise point

then Rule X→Y can be considered a rare rule. Above listed steps forms the "Contradiction check" function defined by the authors of paper [4].

IV. Evaluation

We compare the performance of our model against the performance of existing models as benchmarks. The benchmark models consist of the clustering and association rule generation algorithms used in [4], and an existing rare rule model (RP-Growth) which uses the RP-Tree approach. These comparisons are going to made on:

- Quality of rules generated
- Algorithm runtime

The idea to evaluate the quality of rules is to use a basic classifier. A classifier is trained on a dataset without using any performance tuning. Performance and predictions of the basic classifier contribute as benchmark performance and baseline predictions respectively. Next, these predictions of the classifier are refined using the newly found rare rules. The assumption is that if the rare rules found are indeed informative, then it should capture the underlying pattern and improve the basic classifier performance. This can be applied to any of the class in case of binary classification. Since we are dealing with imbalanced data, it would be interesting to see if it can improve the performance on the minority class as well. Majority and minority classes respectively here refers to the number of instances available for each class. Data sets have been divided into train and test data sets as per hold out method. Hence, these rare rules were only generated using subset of data and being tested on the test data. The choice of classifier and the evaluation strategy is taken from the paper [4] to be able to compare the results. Paper [4] has been evaluated on Random Forest and SVM(Support Vector Machine) classifiers. We added one more neural network based classifier called MLP(Multi layer perceptron). All classifiers are implemented in *sklearn* Library [21].

We re-implemented the DBSCAN clustering algorithm used in paper [4]. Our proposed choice of clustering algorithm, Optics, is taken from the sklearn library [21]. The rare rule generation RP-Growth algorithm implementation is taken from SPFM library [24]. The Association Rule generation Apriori algorithm used in [4] and FP-Growth algorithm is re-implemented by us in Python3. Our ECLAT algorithm and FP-Growth re-implementations are inspired from Python packages [28] and [29] respectively.

We have experimented on the data sets obtained from UCI repository: Breast Cancer, Adult, and Credit Approval as mentioned in Table 1.

TABLE I
DataSet Details

Dataset	Train	Test	Class	Attributes	%MinorityClass
Breast-Cancer	240	46	2	9	.29
Adult	32561	16281	2	14	.24
Credit Approval	598	92	2	15	.44

* All the datasets are taken from UCI Machine Learning Repository.

The F1-Score and confusion matrix will be our performance evaluation measure as it is the most suitable for datasets

with skewed classes. A confusion matrix can capture the effect of using rare rules in the minority class. Since we are aiming to capture the underlying structure in the data, we did not train the classifiers over multiple iterations and avoided dealing with over-fitting problem, which is not the objective of this work. The hypothesis is that for a binary classifier with two class, yes and no, if the trained classifier predicts class no for an instance, the discovered rare rule should be able to predict it as yes based on captured information which was not clear to the underlying model. The same effect can be captured by examining the confusion matrix of the classifier MLP. For a class, an increase in TruePostives and decrease in FalseNegatives is expected after refining existing predictions. We have captured the confusion matrix to show the improvement over minority class only.

Fig. 2 captures the confusion matrix of all the methods on breast cancer dataset. Subfigure 1 in Fig. 2 shows the confusion matrix when MLP classifier is run without using the rare rules generated by any of the methods. When patient does not have cancer, it predicts fairly well, however when patient has cancer it only predicts yes for 5 and does not predict correctly for 9 patients. Subfigure 2 is similar w.r.t. minority class which is class yes. Subfigure 3 shows that in minority class 3, more patients were predicted correctly after the model's predictions were refined based on rare rules captured. Rules in Subfigure 3 were generated using Optics algorithm, and in Subfigure 2 were generated by DBSCAN algorithm. Subfigure 4 shows the results after refining the predictions with rare rules generated by RP-Growth. All of the predictions were made in single class yes. This gives the impression that it worked excellent for minority class but it's no better than a random prediction.

In Fig 4, we show that our approach performs consistently better on all datasets, although in Adult dataset not by much. The scores from all the 3 classifiers were better when the rules generated by our model are used to train classifiers on different datasets. It's interesting to see that the score of classifiers after its predictions are refined by the RP-growth generated rare rules is 50% on all the datasets.

Initial comparison of DBSCAN algorithm and Optics algorithm is interesting to observe. The comparison results of all three datasets are stated in Table 2. On breast cancer dataset, DBSCAN algorithm produced twice the number of noise points than Optics algorithm. However, after applying contradiction check function, number of rare rules generated by DBSCAN was only 3. On the other hand, Optics generated 23 rare rules. For both the algorithms minPts is 10, and eps is 1,0 for DBSCAN algorithm.

In terms of run time we compared the Apriori algorithm to FP-growth and ECLAT algorithm on a range of minimum support from .01 to 0.6. The Rules generated were almost same, however there was a huge difference in run time when run on low minimum support, which is relevant to our goal since rare rules are captured using this category of algorithms with lower values only. Refer to Fig. 4 for the run time graphs. Fig. 4 shows that the FP-Growth algorithm overcomes both Apriori and ECLAT algorithms and performed consistently better.

TABLE II
COMPARISON OF CLUSTERING ALGORITHM

Algorithm	Noise Points	Rare Rules
Breast Cancer Dataset		
DBSCAN	748	3
Optics	386	23
Credit Approval Dataset		
DBSCAN	17269	415
Optics	15291	1650
Adult Dataset		
DBSCAN	11925	0
Optics	7289	4

* Values are based on minpoints =10 for all the datasets.

Table III shows the number of rules generated by the algorithms on the three datasets.

TABLE III
COMPARISON OF ITEMSET GENERATION ALGORITHM

Algorithm	MinSupport	ItemSets
Breast Cancer Dataset		
Apriori	.01	19832
FP-Growth	.01	20146
ECLAT	.01	20146
Adult Dataset		
Apriori	.01	248696
FP-Growth	.01	248696
ECLAT	.01	248696
Credit Approval Dataset		
Apriori	.01	1040738
FP-Growth	.01	1040738
ECLAT	.01	1040738

* This table shows the number of frequent itemsets generated by Apriori, ECLAT and FP-growth algorithms on all 2 datasets. All algorithms generated similar itemsets hence results in similar rules.

Table IV shows the rules generated from RP-growth algorithm and Table V shows the rules generated from our methodology. Clearly there are rare rules generated based on only single element using the RP-growth. However when used with the clustering approach, the rules generated are not straightforward and short. As per the authors of paper [1], one of the characteristics of the rare rules is that they usually contain a large number of attributes, while rare rules generated by RP-growth, for e.g. **menopause = ge40→recurrence = yes**, are extremely short. Intuitively as well predicting a disease only based on 1 attribute is not adequate. The attributes of the rules generated from RP-growth algorithm for credit card approval dataset are anonymised because of data security issues.

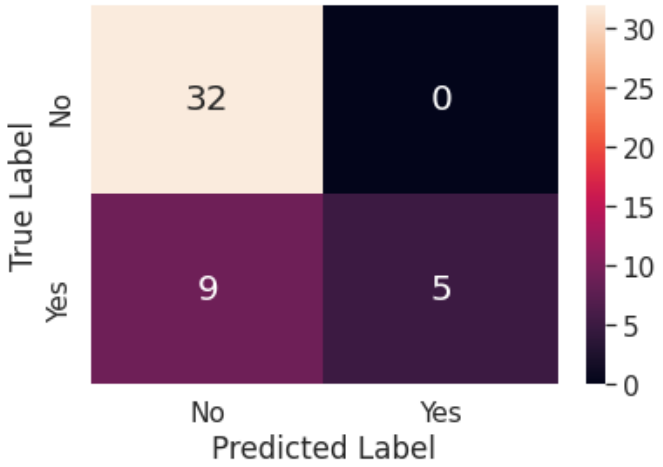

Fig. 1 Confusion Matrix of MLP without Rare Rules and clustering

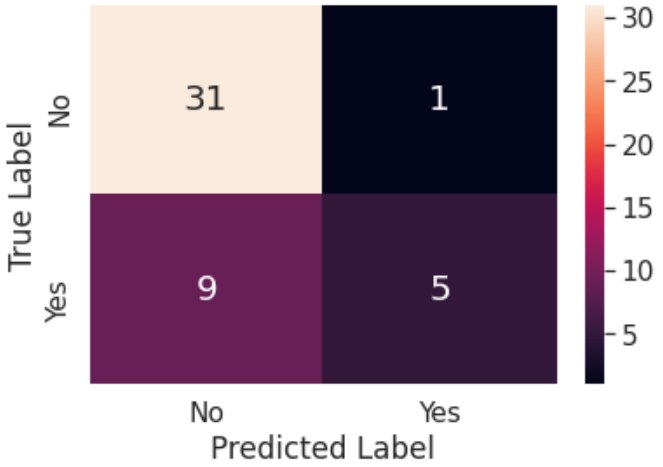

Fig. 2 Confusion Matrix of MLP with Rare Rules and DB-SCAN

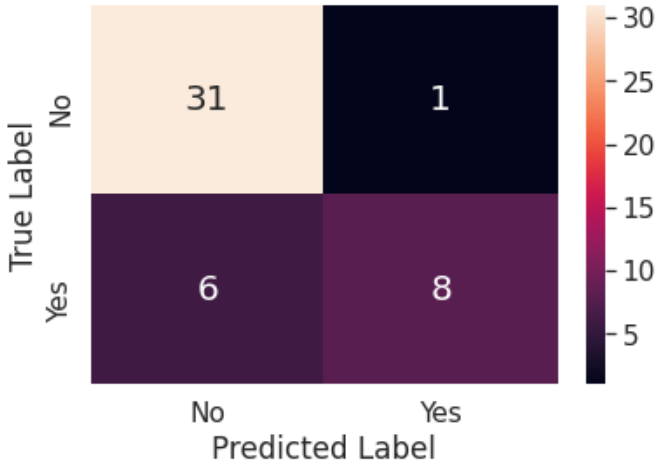

Fig. 3 Confusion Matrix of MLP with Rare Rules and Optics

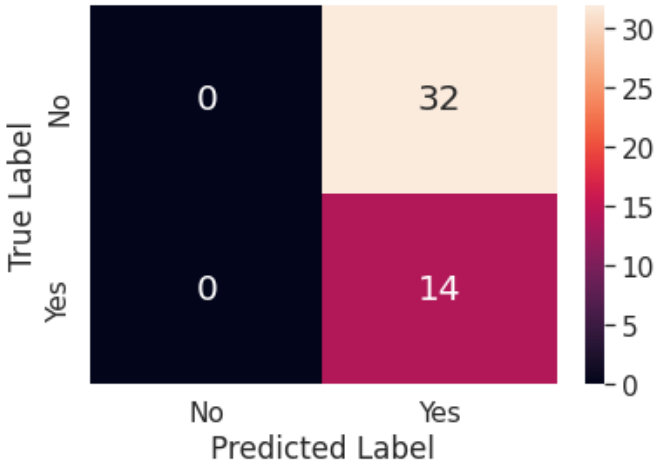

Fig. 4 Confusion Matrix of MLP with Rare Rules by RP-Growth

Fig. 2. Confusion Matrix of MLP classifier on Breast Cancer data-set. Without using any rare rules the classifier is doing good in majority class but not in minority class. Rare rules generated by The DBSCAN method did not improve the performance of minority class, Rare rules generated by RP-growth predicted everything in a single class making it random. Rare rules generated by Optics method did improve the performance of class yes.

Number of total rare rules generated, rare rules generated for minority and majority class is given in the table IV:

TABLE IV
NUMBER OF RULES GENERATED BY RP-GROWTH ALGORITHM

RareRules	MinorityClass	MajorityClass
Breast Cancer Dataset		
304	12	292
Adult Dataset		
1696	208	1492
Credit Approval Dataset		
204	69	135

* This table shows number of rare rules generated by RP-Growth algorithm with 60% support and 10% rare support.

Interpretation of the Frequent and Rare Rules:

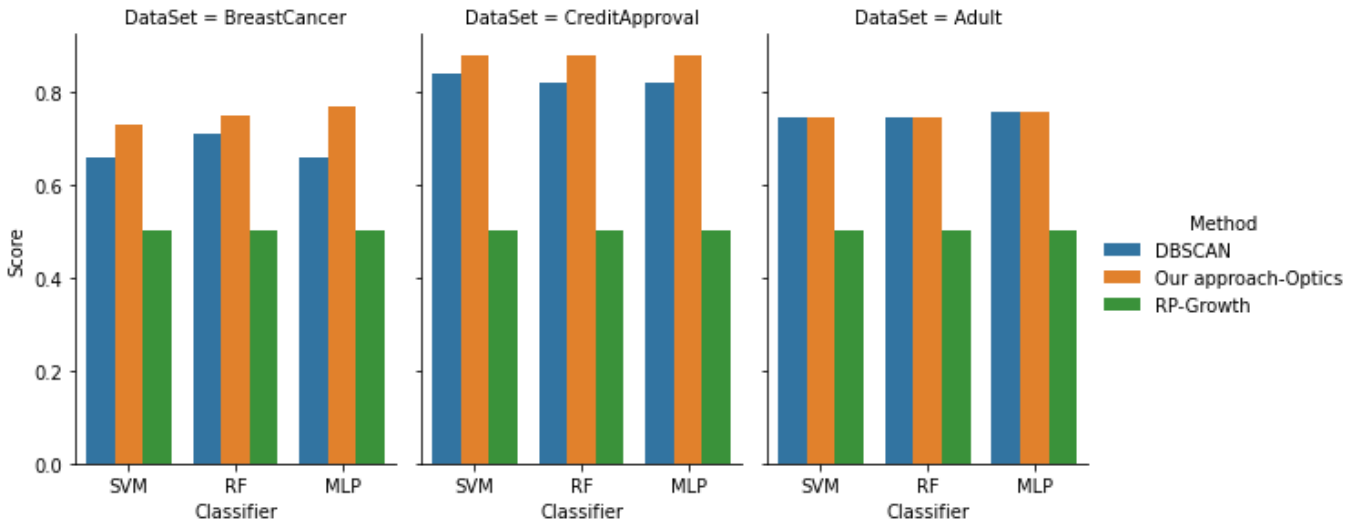

Fig. 3. Performance of the 3 classifiers on Data-sets. F1 -score is compared on each classifier - Multi layer Perceptron, Random Forest, and SVM against Clustering method DBSCAN, our choice Optics and the rare rules generated by RP-Growth Algorithm.

Since the clustering process segregates the frequent rules and rare rules, it's even more reliable and meaningful to be able to compare and look for insights. For example, the rare rule generated by the method based on attribute age=60-69 and deg-malig=2 is **age=60-69, deg-malig=2 →recurrence=no**. The confidence of the rule is 80% which infers that if a woman is of age 60-69 and the degree of malignancy is 2 and there is less chance of recurrence, since it's a binary classification problem it will predict that there would be no recurrence of cancer. However from clustering process we found the rule **age=60-69, inv-nodes=3-5, deg-malig=2 →recurrence=yes** which implies that if in addition to the age and malignancy condition the women also has 3-5 malignant auxiliary lymph nodes, then the recurrence of cancer is yes with 100% confidence. Similarly for the Adult dataset RP-grwoth generated general rules such as **workclass = private →class = >=50k**. If a person works in private sector it infers that income shall be more than 50k. On the other hand clustering approach brings out the rules that **workclass = private, education = bachelors, occupation = Exec-managerial,, relationship = Husband >=50k** which means that if person is a husband, works in private sector at a managerial position, and is a graduate with bachelor's degree, then the income can be more than 50k. The reason for rare rules generated by RP-growth not being specific could be because these are not evaluated based on contradiction against the frequent rule. It's worth noting that clustering approach did not generate rules with single conditions.

V. Discussion

Our choice of algorithms for generating associations rules did consistently better on all chosen data sets. FP-Growth and ECLAT algorithm both did noticeably better than Apriori algorithm when support threshold is kept lowest at .01. FP-Growth did consistently better on all support threshold values, beating ECLAT algorithm on Adult dataset. ECLAT algorithm seems to favour smaller datasets. The choice of clustering algorithm is also supported by the classifier performance, as seen in confusion matrix. Clustering algorithm based approach helped increase the true positives, even for a basic classifier it helped achieve a better F1 score. It did considerably better than the RP-Growth algorithm approach on all datasets. Rules generated by RP-Growth are shorter compared to clustering approaches. RP-Tree takes interesting measures to evaluate the quality of rules which are very similar in nature, and does not evaluate them in the context of performing in combination of a machine learning model. It also generates more number of rare rules for the majority class. However, clustering tends to capture rules which are focused more on the minority class, hence making it much more reliable for imbalanced data sets.

VI. Conclusion and Future Work

In summary, our choice of rule generation and clustering algorithms showed incremental improvements over the baseline methods used by the authors in paper [4]. The baseline paper uses Apriori algorithm for rule generation, we used FP-Growth and ECLAT algorithms, and found substantial improvements in execution time and performance in both cases. FP-growth algorithm beats ECLAT algorithm in run time. The baseline paper used DBSCAN algorithm for clustering, our choice in using the Optics algorithm lead to a reduction in the number of hyperparameters. This translated to an improved performance in terms of the quality of rules. Most importantly, the Optics algorithm approach was entirely independent of the most crucial hyperparameter, eps. Our proposed model with Optics and FP-Growth/ECLAT algorithms worked better than the RP-growth method in terms of quality of rules. Most importantly, we discovered the main drawback of the existing rare rule discovery algorithm RP-Growth, that it does not necessarily generate meaningful rules. Usage of clustering technique in

Rare Rule Generated by RP-Growth
Breast Cancer Dataset
menopause = ge40→recurrence = yes
menopause = premeno→recurrence = yes
menopause = premeno, inv-nodes = 0-2→recurrence = yes
menopause = premeno, node-caps = no→recurrence = yes
menopause = premeno, irradiat = no→recurrence = yes
inv-nodes = 0-2 →recurrence = yes
inv-nodes = 0-2, irradiat = no→recurrence = yes
inv-nodes = 0-2, node-caps = no,irradiat = no→recurrence = yes
inv-nodes = 0-2, node-caps = no→recurrence = yes
node-caps = no→recurrence = yes
irradiat = no→recurrence = yes
node-caps = no,irradiat = no→recurrence = yes
Adult Dataset
workclass = p, race = w→class = g50
workclass = p, race = w, capital-gain = l9999.9, native-country = us →class = g50
relationship = hu, sex = M, capital-loss = l435.6 →class = g50
Credit Approval Dataset
a0 = b, a11 = t→class = yes
a0 = b4, a7 = less-7, a10 = less-17, a11=t→class = yes
a2 = less-7, a7 = less-7, a10 = less-17, a11 = t→class = yes
a7 = less-7, a10 = less-17, a11 = t→class = yes

*These are only the subset of Rare rules containing the attribute menopause and inv-nodes from Breast Cancer dataset, adult datset and credit approval.

TABLE V
RARE RULES GENERATE FROM RP-GROWTH ON DATASETS

Rare Rule Generated by Our method
menopause=ge40, inv-nodes=3-5, node-caps=no→recurrence = yes
age=30-39, menopause=premeno, inv-nodes=0-2,breastquad=left-up→recurrence = yes
age=30-39, menopause=premeno, node-caps=no, breastquad=left-up→recurrence = yes
menopause=ge40, tumor-size=30-34, inv-nodes=3-5→recurrence = yes
menopause=ge40, inv-nodes=6-8, deg-malig=3→recurrence = yes

*These are only the subset of Rare rules containing the attribute menopause from Breast Cancer Dataset.

TABLE VI
RARE RULES GENERATE FROM OUR APPROACH WHEN RECURRENCE OF CANCER IS PREDICTED AS YES

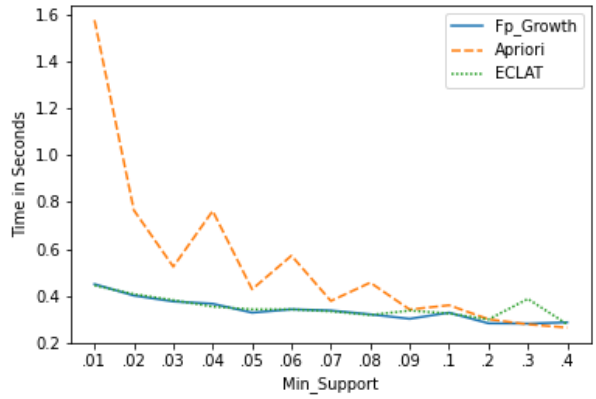

Fig:a Run-time performance on Breast Cancer Data-set

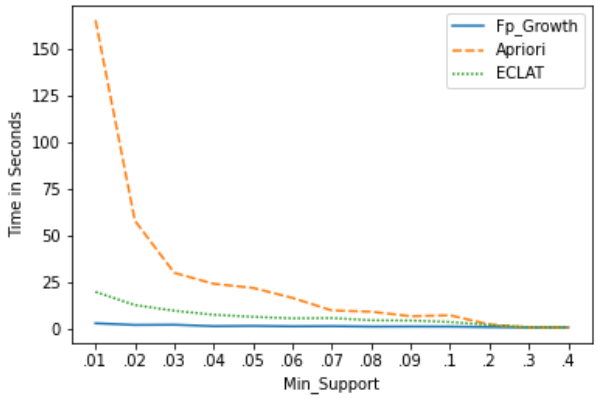

Fig:b Run-time performance on Adult Data-set

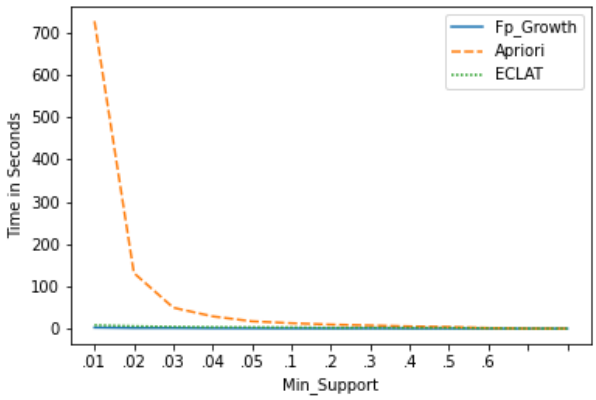

Fig:c Run-time performance on Credit Approval Data-set

Fig. 4. Run time Performance on Data-sets

context of outlier detection in combination of the approach of validating rare rules does bring out more meaningful and informative hidden patterns. Another advantages is that with clustering techniques, user does not need to be an expert in the domain where the problem is being solved, the discovered rare rules are both easier to interpret and compare to the frequent rules.

In terms of future work, further progress over the existing techniques could be achieved by using Auto-encoders. The outlier detection capabilities of an Auto-encoder based system could be compared against a clustering based system to suggest further performance gains. Other approaches specifically designed for imbalanced datasets, which deal with oversampling/undersampling problems with the data, could also be explored. The comparison of their effectiveness when compared to that of clustering based methods should prove enlightening towards a more streamlined approach.

REFERENCES

[1] Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., Mannila, H. (1995). Pruning and Grouping Discovered Association Rules. Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases, 47–52.

[2] Darrab, S., Broneske, D., Saake, G. (2021). Modern Applications and Challenges for Rare Itemset Mining. 11(3). https://doi.org/10.18178/ijmlc.2021.11.3.1037.

[3] Gupta, G. K., Strehl, A., Ghosh, J. (1999). Distance based clustering of association rules. Intelligent Engineering Systems Through Artificial Neural Networks, 9, 759–764.

[4] Bui-Thi, D., Meysman, P., Laukens, K. (2020). Clustering association rules to build beliefs and discover unexpected patterns. Applied Intelligence, 50(6), 1943–1954. https://doi.org/10.1007/s10489-020-01651-1.

[5] Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., He, X. (2020). Generative Adversarial Active Learning for Unsupervised Outlier Detection. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1517–1528. https://doi.org/10.1109/TKDE.2019.2905606.

[6] Prof, L., Kröger, P., Lu, Y. (2017). Knowledge Discovery in Databases II Optional Lecture : Pattern Mining High-D Data Mining Outline • Frequent Itemset Mining – Recap – Relationship with subspace clustering • Rare pattern mining – Relationship with subspace outlier detection.

[7] Pang-Ning Tan, M. S. U., Michael Steinbach, U. of M., Vipin Kumar, U. of M. (2006). Association Analysis: Basic Concepts and Algorithms. Introduction to Data Mining, 238–414. http://www-users.cs.umn.edu/ kumar/dmbook/ch6.pdf.

[8] Koh, Y. S., Pears, R. (2008). Rare association rule mining via transaction clustering. Conferences in Research and Practice in Information Technology Series, 87, 87–94.

[9] Ahmed, M., Barkat Ullah, A. S. S. M. (2018). Infrequent pattern mining in smart healthcare environment using data summarization. Journal of Supercomputing, 74(10), 5041–5059. https://doi.org/10.1007/s11227-018-2376-8

[10] Borah, A., Nath, B. (2019). Rare pattern mining: challenges and future perspectives. Complex Intelligent Systems, 5(1), 1–23. https://doi.org/10.1007/s40747-018-0085-9

[11] Darrab, S., Broneske, D., Saake, G. (2020). RPP algorithm: A method for discovering interesting rare itemsets. In Communications in Computer and Information Science: Vol. 1234 CCIS. Springer Singapore.

[12] Lu, Y., Richter, F., Seidl, T. (2018). Efficient infrequent itemset mining using depth-first and top-down lattice traversal. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10827 LNCS. Springer International Publishing.

[13] Han, J., Pei, J., Yin, Y., Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 8(1), 53–87. https://doi.org/10.1023/B:DAMI.0000005258.31418.83.

[14] Joshi, S., Alsadoon, A., Senanayake, S. M. N. A., Prasad, P. W. C., Yong, S. Y., Elchouemi, A., Vo, T. H. (2020). Pattern Mining Predictor System for Road Accidents. In Communications in Computer and Information Science (Vol. 1287). Springer International Publishing.

[15] Lu, Y., Richter, F., Seidl, T. (2019). LSCMiner: Efficient Low Support Closed Itemsets Mining. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11881 LNCS. Springer International Publishing.

[16] Lu, Y., Richter, F., Seidl, T. (n.d.). Efficient Infrequent Pattern Mining Using Negative Itemset Tree. Springer International Publishing. https://doi.org/10.1007/978-3-030-36617-9.

[17] Van Leeuwen, M., Vreeken, J., Siebes, A. (2006). Compression picks item sets that matter. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4213 LNAI, 585–592.

[18] Tsang, S., Koh, Y. S., Dobbie, G. (2011). RP-tree: Rare pattern tree mining. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6862 LNCS, 277–288.

[19] Kamepalli, S., Bandaru, S. (2019). Weighted Based Frequent and Infrequent Pattern Mining Model for Real-time E-Commerce Databases. Advances in Modelling and Analysis B, 62(2–4), 53–60.

[20] Koh, Y. S., Ravana, S. D. (2016). Unsupervised rare pattern mining: A survey. ACM Transactions on Knowledge Discovery from Data, 10(4). https://doi.org/10.1145/2898359

[21] https://scikitlearn.org

[22] Garg, Kanwal Kumar, Deepak. (2013). Comparing the Performance of Frequent Pattern Mining Algorithms. International Journal of Computer Applications. 69. 21-28. 10.5120/12129-8502.

[23] Sidney Tsang, Yun Sing Koh, Gillian Dobbie, RP-Tree: Rare Pattern Tree Mining, International Conference of Data Warehousing and Knowledge Discovery, 277-288 (2011).

[24] http://www.philippe-fournier-viger.com/spmf/.

[25] Yun Sing Koh and Nathan Rountree. 2005. Finding sporadic rules using apriori-inverse. In PAKDD (Lecture Notes in Computer Science), Tu Bao Ho, David Cheung, and Huan Liu (Eds.), Vol. 3518. Springer, Berlin, 97–106.

[26] Bhatt, U.Y., Patel P.A.: An effective approach to mine rare items using Maximum Constraint. In: Intelligent Systems and Control (ISCO), IEEE, pp. 1–6 (2015)

[27] Weiss, G.M.: Mining with rarity: a unifying framework. In: ACM SIGKDD Explorations Newsletter, pp. 7–19 (2004)

[28] https://pypi.org/project/pyECLAT/

[29] https://pypi.org/project/pyfpgrowth/