



OPECUR: An Enhanced Clustering-Based Model for Discovering Unexpected Rules

Sadeq Darrab¹(✉), Priyamvada Bhardwaj¹, David Broneske²,
and Gunter Saake¹

¹ Otto von Guericke University, Magdeburg, Germany

{sadeq.darrab, Gunter.Saake}@ovgu.de, priyamvada.bhardwaj@st.ovgu.de

² German Centre for Higher Education Research and Science Studies,
Hannover, Germany
broneske@dzhw.eu

Abstract. Rare association rules gained importance since they are widely used in critical real-life domains, such as medicine, fraud detection, malware attacks, recommender systems, and weather forecasting. In order to extract actionable rules that can be used in real-life scenarios, user confidence and an easy-to-use model with as few as possible tuning knobs are required. On top of this, an inherent imbalance of the data (e.g., in the medical domain, there are fewer ill people compared to healthy people) poses a severe challenge, which complicates the finding of rare patterns. Recently, an unsupervised clustering model was proposed to discover interesting rare rules. However, the performance of this model degrades in terms of time and accuracy. In this paper, we propose an efficient model to recover interesting rare rules. In this model, we employ machine learning-based classifiers to assess the performance of the generated rules. To evaluate the proposed model, we experiment with three real-life medical datasets. The experimental results show that our model outperforms the state-of-the-art model in terms of time and accuracy. Furthermore, we generate more accurate results, which means that the user is only confronted with the most important and compact rules. This reduces a user's effort on postprocessing of associating rules significantly.

1 Introduction

Frequent itemset mining is the process of finding a set of items (known as pattern or itemset) that co-occur in a given dataset. If a strong relation between items can be inferred, they form an association rule. An inferred association rule is represented as $X \rightarrow Y$, where X and Y are a set of itemsets. For example, in market basket analysis, a rule $\text{milk} \rightarrow \text{bread}$ represents that if a customer buys milk, the customer also tends to buy bread as well. Generating association rules from a dataset is called association rule mining (ARM). ARM gains special importance due to its advantages in many domains, such as market basket analysis, medical diagnosis, and web usage mining [1, 2]. The primary goal of ARM

is to produce rules that are sensible and beneficial to the user. However, a vital shortcoming of ARM is that it usually generates many rules (i.e. for an itemset with d items, there are $2^d - 1$ rules that can be generated [20]). Extracting such a large number of association rules degrades the performance, making it challenging for a decision-maker to analyze the produced rules. Hence, the resulting rules require extensive downstream analysis. To overcome these shortcomings, it is practical to produce fewer and more meaningful rules for the user.

One of the most frequently used techniques for the simplification of rules is called clustering association rules [10, 15, 19]. Clustering association rules generate a condensed representation of association rules and, thus, less rules compared to those generated by conventional methods. Therefore, extracting compact and understandable rules leads to more accurate classifiers and produces a manageable number of rules for the end-user for further investigation. There is a plethora of work proposed for clustering association rules [6, 10, 15]. Most of these methods focus on generating rules from frequent patterns. However, the generated rules represent common events and are, thus, usually expected and unsurprising. Therefore, analyzing the data at hand for more meaningful and unexpected patterns is desirable to reveal unknown knowledge in many real-life applications. To this end, discovering unexpected association rules (also known as rare itemset mining) came into focus [8]. To reveal such knowledge, the state of the art is to use an unsupervised machine learning method, DBSCAN-based clustering [5], to extract unexpected association rules. Although this method finds unexpected rules, it comes with the following shortcomings. First, it is costly in terms of time and memory since it utilizes the Apriori algorithm to generate the complete set of rules [18]. Second, adjusting the parameters properly (i.e., ϵ , $minPts$ for DBSCAN) is challenging since DBSCAN as a clustering-based method may fail to identify unexpected rules when choosing improper parameters. Lastly, the F1-score of a classifier was utilized in the DBSCAN-based model to evaluate the quality of generated rules. However, the reported score was observed only for a single class (i.e., minority class), leading to ambiguity when used on data with class imbalances [12]. For instance, an improved F1-score of a class might also mean that the classifier improved its prediction at the cost of other class scores in imbalanced data (i.e., predicting that everyone is ill).

Due to these shortcomings, designing an efficient model to address these three gaps is still a critical research problem. This paper deals with these limitations by developing an efficient model, *OPTICS*-based clustering of *ECLAT*-generated *Unexpected Rules* (*OPECUR*), to find unexpected association rules efficiently and accurately. The main contribution of this paper is as follows.

- To generate the whole set of association rules, we employ FP-Growth and ECLAT algorithms to extract the complete set of association rules efficiently.
- For the clustering task, we utilize a density-based clustering algorithm, OPTICS, to avoid the problem of extensive hyperparameter tuning. OPTICS can also detect clusters of different densities efficiently leading more precise unexpected association rules.

- We utilize three different machine learning classifiers for the classification task in the evaluation process: Support vector machines (SVM), Random forest (RF), and Neural network-based multi-layer perceptron (MLP).
- The experimental results show that the F1-score and area under the curve (AUC) of our OPECUR model are constantly better than the state-of-the-art DBSCAN-based model. Furthermore, our experimental results show that our model recovers association rules much faster than the state of the art and generates more interesting unexpected association rules.

The remainder of this paper is structured as follows. In Sect. 2, we provide background on association rules and present related work in Sect. 3. In Sect. 4, our proposed method adopted in the paper is explained. Our evaluation strategy, performance criteria, and variations are discussed in Sect. 5. We end the paper with a conclusion in Sect. 6.

2 Background

This section briefly presents important fundamental knowledge for discovering unexpected rules. One primary data mining task is association rule mining aiming at discovering hidden valuable knowledge from a dataset. The main step of ARM is to find a set of items (known as patterns or itemset) that co-occur in the dataset. Hence, association rule mining requires the following two main steps:

1. The complete set of interesting patterns is extracted.
2. The association rules are generated from the set of desired patterns that have been mined in the first step.

We describe the problem of association rule mining and its related concepts, such as frequent and unexpected (rare) patterns with the following definitions. Let DB be a dataset of m transactions such as $DB = \{T_1, T_2, \dots, T_m\}$ and I represents the set of unique n items in DB , $I = \{i_1, i_2, \dots, i_n\}$. Some common definitions are defined below to illustrate the concept of association rule mining.

Definition 1 (Support of a pattern). Given a pattern X , its support (Sup) is defined as the number of transactions in DB containing the pattern X such that $Sup(X) = \frac{Sup(X)}{|DB|}$, where $|DB|$ represents the size of the dataset, DB .

Definition 2 (Frequent and Rare Patterns). A pattern X whose support satisfies a user-specified support threshold, $minSup$, is called frequent pattern such that $sup(X) \geq minSup$. In contrast, the pattern X that does not satisfy $minSup$ is called a rare pattern.

Definition 3 (Strong Association Rule). An association rule $X \rightarrow Y$ is called strong if its support and confidence measures satisfy a specified minimum support threshold ($minSup$) and minimum confidence threshold ($minConf$).

Definition 4 (Unexpected Association Rule [16]). A rule $X \rightarrow Y$ is an unexpected rule w.r.t. a known rule $A \rightarrow B$ if the following conditions hold:

1. The antecedents of rules (i.e., A and X) statistically hold on DB . Besides, they are similar according to a statistical measure (i.e., cosine similarity exceeds a given threshold).
2. The consequences of rules (i.e., B and Y) contradict each other.

In this paper, we focus on discovering unexpected association rules that satisfy Definitions 3 and 4.

3 Related Work

Association Rule Mining (ARM) identifies interesting relations among patterns in a dataset. ARM gains vital importance since it helps to turn data into useful knowledge in various domains such as fraud detection, disease diagnosis [3], and road traffic accident prediction [13]. There is a plethora of methods presented to recover the complete set of association rules [1, 11, 14, 21]. The Apriori and FP-Growth algorithm [18] and their extended versions are well-known methods that are extensively used for ARM. These methods work well for recovering the complete set of association rules, but generate too many rules leading to a scalability issue. Besides, producing a massive amount of rules makes it challenging for a decision-maker to analyze the produced rules.

To this end, clustering techniques can produce fewer, more concise rules in many models by grouping similar rules into clusters [6, 10, 15, 19]. In [19], a method has been presented to prune undesired association rules and group similar rules into clusters. They use the concept of a rule cover to prune uninteresting rules and then apply clustering to get fewer, more concise rules. Further methods are the “BitOp” algorithm [15], or the usage of conditional market-basket difference (CMBP) and conditional market-basket log-likelihood (CMBL) to group association rules by utilizing agglomerative clustering [10]. The latter generates association rules that are grouped based on a normalized distance metric. In [6], Dhahi et al. group association rules into disjoint clusters using K-means. All these approaches use frequent itemsets.

Rare rules can capture hidden patterns in data that may be important in rare disease diagnosis and credit card fraud detection. Instead of generating the whole set of rules, there are several algorithms that generate rare rules [7, 8]. Although these methods produce interesting rare association rules, they have some limitations, such as 1) finding the rare items before the mining process is challenging and 2) they generate many rare rules. To address these shortcomings, a clustering-based model has been proposed to mine unexpected rare rules by utilizing DBSCAN to cluster rare rules [5]. After clustering association rules, the DBSCAN based model checks whether rules are noise rules or unexpected rules based on a contradiction check. Although this model finds unexpected rules, it comes with notable limitations. First, its performance is not optimal since it employs the Apriori algorithm to mine the rules. Second, the interesting

unexpected rules may be missed since the DBSCAN algorithm suffers from the problem of not detecting nested cluster structures [4]. Lastly, the hyperparameters of the DBSCAN algorithm are crucial to the clustering results and greatly influence the overall results. Hence, setting up proper hyperparameters is challenging. In this paper, we propose a model to address these challenges as in the following section.

4 Proposed Method: OPECUR Model

This paper proposed a clustering model, the OPECUR model, to discover valuable unexpected rare association rules. Figure 1 shows the workflow of the model to obtain the unexpected rare rules. The model first derives all the association rules from a dataset. The minimum support is kept low for generating the rules so that the chance of missing important rare rules can be eliminated. Furthermore, all the association rules are converted to feature vectors using the correlation of items/attributes present in a dataset. Next, the contradiction check function, which we describe at the end of the clustering algorithm section, is applied to the noise points marked by the clustering algorithm to determine the final unexpected association rules. Our proposed model consists of two main phases (i.e., Association rules generation and clustering process see Fig. 1). In the first phase, the complete set of association rules is generated by utilizing efficient methods such as FP-growth and ECLAT. The second step explains the process of generating unexpected rules by utilizing the OPTICS clustering technique [4].

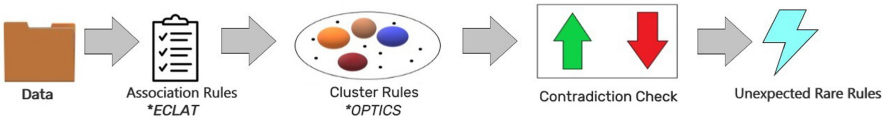


Fig. 1. OPECUR workflow to generate unexpected rare rules

4.1 Generating Association Rule

The state of the art approach [5] generates the whole set of association rules using the Apriori algorithm. However, Apriori has the shortcoming of a huge amount of candidates generated and it has to take multiple passes over the dataset. Our proposed model, OPECUR, addresses these limitations by employing faster, more scalable, and more efficient algorithms (i.e., FP-Growth and ECLAT algorithms) to generate association rules. The FP-Growth [11] algorithm uses a tree structure called FP-Tree to hold all necessary information for the mining process without a candidate and test generation process. FP-Growth scans a dataset at most twice. In the first scan, the FP-Tree is constructed by adding transactions (i.e., after removing useless items) one by one. After the tree

is built, the mining process starts by employing a divide-and-conquer approach. The ECLAT algorithm [21] uses a depth-first search approach to generate the whole set of patterns. ECLAT works vertically and does not require multiple passes over the data. Moreover, it applies intersections to find the support count of the generated patterns.

4.2 Clustering Algorithm

Our proposed model groups similar association rules together using a density-based algorithm *OPTICS* [4]. The OPECUR 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. Furthermore, the concept of reachability distance reduces the necessity of providing an *eps* value since it is automatically calculated based on *minPts*. This helps in dealing with different densities within a cluster and discovering nested clusters. As it can be seen in Fig. 2 and Fig. 3, the OPECUR model generates more clusters. Since *eps* is calculated automatically in the OPTICS algorithm, it is able to discover the clusters which are not available with a given value of *eps* like in DBSCAN. Notably, red points in Fig. 2 and Fig. 3 are noise points. To recover unexpected rules, we investigate the noise rules that may contain interesting unexpected rules. To identify whether a noise rule is an unexpected rule or not, we use a similar approach as used in [5] with the following difference. We first run the OPECUR algorithm and then run the contradiction check function on the potential noise rules to find interesting unexpected rules. It is different from the method used in [5] as they employ the contradiction check property as a part of the clustering algorithm. Thus, the contradiction check function affects the results since adding multiple parameters to tune at once making it harder and undesirable. In contrast, OPECUR model generates noise rules for a *minPts* value and then runs the contradiction check function. Hence, our model reduces the chances of missing the most informative unexpected rules.

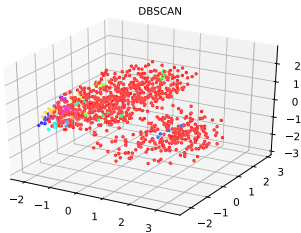


Fig. 2. Clusters on breast cancer dataset using DBSCAN

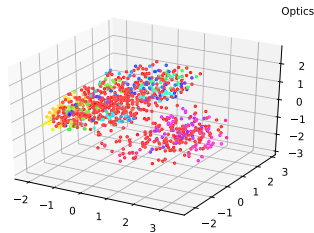


Fig. 3. Clusters on breast cancer dataset using OPECUR

As per the contradiction check function, if there exists two rules $X \rightarrow Y$ and $X' \rightarrow Y'$ the following conditions are checked:

1. $Y \neq Y'$
2. Cosine similarity of X and X' is high
3. Confidence of both rules is high
4. Rule $X \rightarrow Y$ is a noise point

The noise rule $X \rightarrow Y$ is compared to the cluster rules and it is marked as an unexpected rule in contradiction to a frequent rule $X' \rightarrow Y'$ if all of the above conditions are satisfied.

5 Experimental Evaluation

In this section, our proposed model’s performance is compared with the state-of-the-art model introduced in [5]. In the following, we introduce the evaluation setup and then explain the results of our conducted experiments.

5.1 Experimental Setup

We compare our model, OPECUR, with the state-of-the-art model introduced in [5] by utilizing three real-life medical datasets. The datasets are Breast Cancer, Cleveland Heart Disease, and Hepatitis obtained from the UCI repository [9]. In Table 1, we summarize the main characteristics of the datasets. The Cleveland heart disease dataset has four target variables for prediction, 0 (absence) and 1, 2, 3, 4 (presence)¹. Hepatitis and Cleveland datasets consist of real-valued and categorical attributes. Breast Cancer and Hepatitis datasets have a non-uniform distribution of instances 29% and 25% respectively in the minority classes. The skewed data distribution of the instances in the classes poses a significant challenge in finding fewer representative and meaningful rare rules from the datasets.

To test the efficiency and effectiveness of our new approach, we execute three experiments. In the first experiment, we compare our proposed OPECUR model with the DBSCAN-based model in terms of time required to generate the complete set of rules in order to assess its scalability. In order to test the quality of our clustering, the second experiment compares found clustered association rules from our OPECUR with the DBSCAN-based model. Besides, we evaluate the quality of generated rare rules from our clustering algorithm OPTICS with the DBSCAN algorithm by training a classifier with the representation of the rules. This will assess whether the generated rules are meaningful for decision making (i.e., identifying ill or healthy people). All the experiments ran on Google Colab with a limited 12 GB of RAM to test the scalability and easy-to-use nature of the models. The results are presented in the following subsections.

¹ It is converted to binary classification, 0 (absence), and 1 (presence) for simplicity.

Table 1. Dataset details

Dataset	Instances	Class	Attributes	%MinorityClass
Breast cancer	286	2	9	.29
Hepatitis	155	2	20	.25
Cleveland	303	2	14	.44

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

5.2 Experiment 1: Execution Time Comparison

This section evaluates the execution time of our proposed OPECUR model with the state-of-the-art model using DBSCAN for generating the complete set of association rules. While the DBSCAN-based model utilizes Apriori, we also test here FP-Growth and ECLAT algorithms to generate the whole set of rules. For all experiments, the minSup is varied from 0.01 to 0.04. These algorithms generate the exact same set of rules.

Figure 4, 5 and 6 shows the performance of the algorithms on all datasets. As it can be seen, FP-Growth and ECLAT consume less time compared to the Apriori algorithm. This is not totally surprising, because FP-Growth and ECLAT algorithm outperformed the Apriori algorithm for frequent itemset mining in various studies [18]. However, compared to higher minSup values in frequent pattern mining, it can be seen that even with low minSup, FP-Growth and ECLAT algorithms recover patterns faster. This is crucial in generating the complete set of rules, including interesting rare ones. Hence, our algorithm relies on the usage of ECLAT algorithm to generate the complete set of rules since it shows better performance when mining itemsets with a low support threshold.

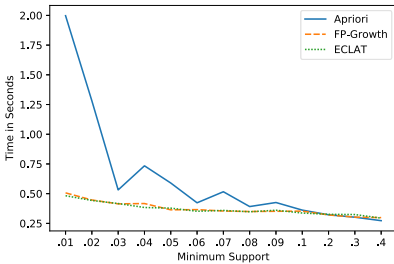


Fig. 4. Runtime performance on breast cancer dataset

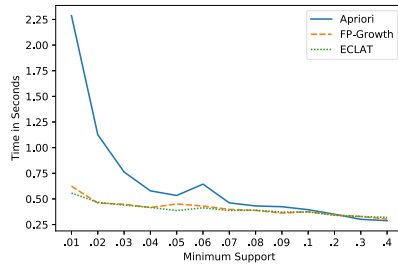


Fig. 5. Runtime performance on cleveland heart disease dataset

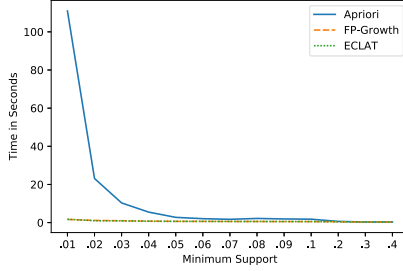


Fig. 6. Runtime performance on hepatitis dataset

5.3 Experiment 2: Clustering Process Comparison

In this experiment, we compare the results of the clustering process using our proposed model, OPECUR, with the state-of-the-art DBSCAN-based model. To reach comparable results, we set the parameters for the whole experiment similar to the compared model, DBSCAN-based [5] (i.e., minPts to 10, delta1 and delta2 of the contradiction check to 1 and -1 , respectively). Table 2 shows the post clustering statistics from the OPECUR and DBSCAN-based algorithms. The results show that our model generates more unexpected rare rules than those generated by the DBSCAN algorithm. Hence, our model successfully recovers the interesting rare rules, whereas the DBSCAN model misses these unexpected rare rules. This is because the OPECUR model automatically calculates the minEps parameters and can find more clusters in the data than the DBSCAN algorithm, resulting in lesser noise rules. Hence, experimental results show that our proposed model can find more interesting unexpected rare rules that may be missed by the state-of-the-art DBSCAN-based model. However, the question whether these are meaningful is still open and will be evaluated.

To evaluate the quality of unexpected rare rules generated by our proposed OPECUR model, we adopted a similar evaluation strategy as in [5]. At its heart, they measure the impact of rare rules through using machine learning-based classifiers. We adhere to their assumption that if a rule improves the performance of a classifier by refining its decision boundary, then the rule indeed is beneficial. Beyond these assumptions, we change the evaluation procedure in the following way: As the first revision, in the evaluation procedure, we perform a 3-fold cross validation on the datasets to train and test the model instead of the independent hold-out method. Hence, the rules in two folds are used as a training dataset, whereas the remaining third fold is used as a test dataset. The process is repeated until each fold has served as an independent test set to be evaluated. We report the average scores of the models after cross validation for a better estimate of the generalization error. For the classification task, we utilize Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) classifiers

implemented in Sklearn library [17] with default classifier parameters. We report F1 and AUC metrics of the classifiers for the evaluation metric. F1 score considers precision and recall of the classifier, and AUC (Area under the curve) gives a better idea of the balance between true-positive and false-positive rates.

Figure 7 and Fig. 8 show the performance of our proposed model, OPECUR, and the state-of-the-art model, DBSCAN-based, on all three datasets. For both measures, our proposed model, OPECUR, shows consistently better performance compared to the state-of-the-art DBSCAN-based model [5], on all three datasets. This is due to the fact that OPECUR can find more clusters and generate fewer valid noise rules that pass the contradiction check. This is a clear advantage compared to the DBSCAN model, which only finds a limited number of the actual amount of clusters. Hence, rules that should ideally be marked as noise will still be assigned to a cluster by DBSCAN and, thus, be considered as a frequent rule. Thus, our proposed model is able to generate more unexpected rare rules and gains better performance in terms of F1 and AUC score.

Table 2. Comparison of clustering algorithms

Algorithm	Noise	Rare	Clusters
Breast Cancer Dataset			
DBSCAN	748	3	20
OPTICS	386	23	54
Cleveland Dataset			
DBSCAN	17269	415	4
OPTICS	15291	1650	17
Hepatitis Dataset			
DBSCAN	11925	0	3
OPTICS	7289	4	10

*Values are based on minpoints = 10 for all the datasets. The table describes the number of clusters, noise rules and rare rules generated by each algorithm on all the datasets.

Notably, our OPECUR methods shows the highest performance for the Hepatitis dataset. This is due to the fact that it contains more attributes, which benefits the identification of correlations in these attributes. Furthermore, our in-depth analysis of the rules and clusters shows that this dataset is a challenge for DBSCAN as nested clusters in the Hepatitis dataset have not been identified.

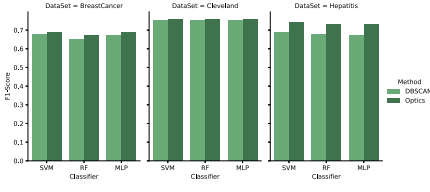


Fig. 7. F1 Score (F1-Score and AUC-Score of the three classifiers on the datasets when using the rules from DBSCAN or OPECUR (SVM = Support Vector Machine, RF = Random Forest, MLP = Multi-Layer Perceptron).)

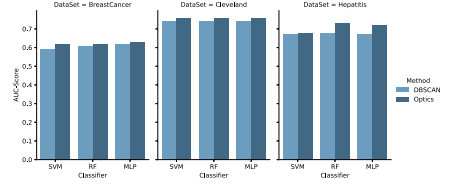


Fig. 8. AUC Score (F1-Score and AUC-Score of the three classifiers on the datasets when using the rules from DBSCAN or OPECUR (SVM = Support Vector Machine, RF = Random Forest, MLP = Multi-Layer Perceptron).)

5.4 Experiment 3: Evaluation of Unexpected Rules

In this section, we evaluate the generated unexpected rules by two criteria. Firstly, we evaluated them by utilizing the contradiction check approach that used in the compared model [5] with same parameters. According to this property, our model generates more interesting unexpected association rules. For instance, a rule ‘age = 50–59’, ‘breast = left’, ‘deg-malig = 3’, ‘irradiat = no’, ‘menopause = ge40’, ‘tumor size = 30–34’ → class = yes is marked as unexpected rule by OPECUR. The rule is in contradiction to the following two rules:

- ‘age = 50–59’, ‘breast = left’, ‘irradiat = no’, ‘menopause = ge40’ → class = no.
- ‘breast = left’, ‘irradiat = no’, ‘menopause = ge40’ → class = no.

Secondly, we compare the unexpected rules generated by the rare-pre-post-order algorithm, RPP [7]. RPP algorithm is proposed recently to discover the whole set of rare rules. The results show that the a set of unexpected association rules generated by our proposed model, OPECUR, is a subset to the rare rules generated by the RPP algorithm, providing meaningful and concise information.

For example, a rare rule generated by RPP algorithm is the following: ‘tumor size = 30–34’, ‘inv-nodes = 3–5’, ‘node-caps = no’, ‘menopause = ge40’, ‘deg-malig = 3’, ‘irradiat = no’ → class = yes. The rule states that if a patient has a tumor of diameter 30–34, 3–5 auxiliary lymph nodes, node caps is no, degree of malignancy of the node is 3, and patient is menopausal without any history of radiation therapy taken, then the chances of recurrence of cancer are higher. In comparison, our model marks the following rule as rare: ‘menopause = ge40’, ‘inv-nodes = 3–5’, ‘node-caps = no’, ‘irradiat = no’ → class= yes. Hence, the same information generated by our model in this rules is also generated by the RPP algorithm. Thus, the results show that the unexpected association rules generated by our model are informative and meaningful.

6 Conclusion

Discovering unexpected (rare) rules gains attention since it reveals interesting hidden knowledge from various domains such as medical diagnosis, fault detection, and fraud detection. However, the state-of-the-art model's performance degrades when generating unexpected association rules and fails to generate the whole set of unexpected association rules. To address these limitations and discover unexpected rules, we designed a clustering-based model, OPECUR, to recover unexpected association rules from real datasets. The proposed model addresses the state of the art's shortcomings to efficiently generate the complete set of rules. Besides, our proposed model generates more interesting unexpected rules due to its capability to set parameters automatically during the clustering process. Hence, our model can find more clusters, resulting in more unexpected rules that helped improve the decision boundary of a machine learning classifier. The discovered unexpected rules are assessed based on several criteria: contradiction check property, classifiers' performance, and the RPP algorithm's results. The experimental results show that our model is scalable and generates more trustworthy unexpected rules. In the future, we intend to investigate this model further on streaming data.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB, pp. 487–499. Morgan Kaufmann Publishers Inc. (1994)
2. Ahmed, M., Barkat Ullah, A.S.S.M.: Infrequent pattern mining in smart healthcare environment using data summarization. *J. Supercomput.* **74**(10), 5041–5059 (2018). <https://doi.org/10.1007/s11227-018-2376-8>
3. Altaf, W., Shahbaz, M., Guergachi, A.: Applications of association rule mining in health informatics: a survey. *Artif. Intell. Rev.* **47**(3), 313–340 (2016). <https://doi.org/10.1007/s10462-016-9483-9>
4. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: ordering points to identify the clustering structure. *SIGMOD Rec.* **28**(2), 49–60 (1999). <https://doi.org/10.1145/304181.304187>
5. Bui-Thi, D., Meysman, P., Laukens, K.: Clustering association rules to build beliefs and discover unexpected patterns. *Appl. Intell.* **50**(6), 1943–1954 (2020). <https://doi.org/10.1007/s10489-020-01651-1>
6. Dahbi, A., Mouhir, M., Balouki, Y., Gadi, T.: Classification of association rules based on K-means algorithm. In: *CiST*, pp. 300–305. IEEE (2016). <https://doi.org/10.1109/CIST.2016.7805061>
7. Darrab, S., Broneske, D., Saake, G.: RPP algorithm: a method for discovering interesting rare itemsets. In: Tan, Y., Shi, Y., Tuba, M. (eds.) *DMBD 2020. CCIS*, vol. 1234, pp. 14–25. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-7205-0_2
8. Darrab, S., Broneske, D., Saake, G.: Modern applications and challenges for rare itemset mining. *IJMLC* **11**(3), 208–218 (2021). <https://doi.org/10.18178/ijmlc.2021.11.3.1037>
9. Dua, D., Graff, C.: UCI machine learning repository (2017). <http://archive.ics.uci.edu/ml>

10. Gupta, G.K., Strehl, A., Ghosh, J.: Distance based clustering of association rules. *Intell. Eng. Syst. Artif. Neural Networks* **9**, 759–764 (1999)
11. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Min. Knowl. Discov.* **8**(1), 53–87 (2004). <https://doi.org/10.1023/B:DAMI.0000005258.31418.83>
12. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data-recommendations for the use of performance metrics. In: *International Conference Affect Computing Intelligence Interact Workshops*, pp. 245–251 (2013)
13. Joshi, S., et al.: Pattern mining predictor system for road accidents. In: Hernes, M., Wojtkiewicz, K., Szczerbicki, E. (eds.) *ICCCI 2020. CCIS*, vol. 1287, pp. 605–615. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63119-2_49
14. Kamepalli, S., Bandaru, S.: Weighted based frequent and infrequent pattern mining model for real-time e-commerce databases. *Adv. Model. Anal. B* **62**(2–4), 53–60 (2019). <https://doi.org/10.18280/ama.b.622-404>
15. Lent, B., Swami, A., Widom, J.: Clustering association rules. In: *ICDE*, pp. 220–231 (1997). <https://doi.org/10.1109/icde.1997.581756>
16. Padmanabhan, B.: A belief-driven method for discovering unexpected patterns. In: *KDD*, pp. 94–100. AAAI Press (1998)
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: machine learning in Python. *JMLR* **12**, 2825–2830 (2011)
18. Singh, A.K., Kumar, A., Maurya, A.K.: An empirical analysis and comparison of apriori and FP-growth algorithm for frequent pattern mining. *ICACCCT*, pp. 1599–1602 (2015). <https://doi.org/10.1109/ICACCCT.2014.7019377>
19. Toivonen, H., Klemettinen, M., Ronkainen, P., Hättönen, K., Mannila, H.: Pruning and Grouping Discovered Association Rules. In: *Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pp. 47–52 (1995)
20. Zaki, M.J.: Generating non-redundant association rules. In: *SIGKDD*, pp. 34–43. Association for Computing Machinery (2000). <https://doi.org/10.1145/347090.347101>
21. Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: New algorithms for fast discovery of association rules. In: *KDD*. AAAI Press (1997)