EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Top 'N' Variant Random Forest Model for High Utility Itemsets Recommendation

Pazhaniraja N[1], Sountharrajan S[1,*], Suganya E[2] and Karthiga M[3]

[1]Department of Computing Science and Engineering, VIT Bhopal University, Sehore, MP, India-466114
[2]Anna University, Chennai, India
[3]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu 638401, India.

Abstract

High-utility based itemset mining is the advancement of recurrent pattern mining that discovers occurrence of frequent transactions from a huge database. The issues in frequent pattern mining involve the elimination of quantities purchased by the customers and cost of purchased product. This can be resolved by high utility itemset mining which includes quantities and profit of the products in the transactions. The conventional association rule mining algorithms results in huge memory consumption due to the complexity in pruning the search space. In this paper, machine learning based high-utility itemset mining is applied to predict next order in an online grocery store depending on the transactions. The overall goal is to enhance the business profitability by stocking the high utility items in market. The Top 'N' variant Random Forest model is proposed to recommend the high utility itemsets, thereby predicting the reordered/next ordered items. The model is evaluated using Instacart market dataset to measure accuracy, precision and recall.

Keywords: High Utility Itemset, Random forest, machine learning, association mining, frequent itemsets, feature selection.

Received on 05 November 2020, accepted on 15 December 2020, published on 25 January 2021

doi: 10.4108/eai.25-1-2021.168225

*Corresponding author. Email: Sountharrajan@gmail.com

1. Introduction

Decision making is the significant part of a profitable business strategy that gains insight from the real time transactional databases. The Customer Relationship Management (CRM) is the process of gaining the details of the customers and their preferred products using which their behavioural patterns are framed (1). The data mining technique greatly assists in mining the patterns from huge real-world databases, thereby enhancing the decision building process. Transactional status of customers is utilized to recognize the purchase pattern and to improve the business profitability (2). Association Rule Mining (ARM) that helps in identifying the frequently accessed itemsets from the database is termed as Frequent Itemset Mining (FRM) (3). The ARM has its wide applications in market basket analysis, recommendation systems, bioinformatics, network analysis, customer reviews, intrusion detection, image classification, and so on. The conventional ARM algorithms include FP-growth tree and Apriori algorithm (4) (5) used to discover the frequently used itemsets. In general, the mining algorithms are classified as constrained based algorithms, tree-based algorithms, projection-based algorithms and apriori based algorithms. The customer behaviour can be recognized by mapping the association between the items purchased that helps in the promotion of the products, thereby increasing the profit. The itemset mining involves the discovery rare, frequent and correlated itemset. Apart from the conventional way of association rule mining, the machine learning based high utility itemset mining also plays a significant role in decision making. The candidate key generation and rule based frequent itemsets are integrated in the classification-based association rule mining. The speed and easy understanding of decision tree

grabbed the attention of researchers to predict the frequently accessed items from the dataset (6). Let the transaction database (TD) has n number of unique items and't' transactions which is represented as follows. $TD = \{td^1, td^2, \dots td^t\}$and itemset $I = \{i^1, i^2, \dots i^n\}$. All the transactions are enclosed in the itemset $\mathrm{I}(t \in I)$. Each transaction is represented with a sole identifier known as Transaction-ID. A sample of transaction dataset with five transactions is shown in Table 1.

Table 1. Transactions' sample in a dataset

TID	Transactions
T_1	$\{a_1,c_1,d_1\}$
T_2	$\{b_1,c_1,e_1\}$
T_3	$\{a_1,b_1,c_1,e_1\}$
T_4	$\{b_1,e_1\}$
T_5	$\{a_1,b_1,c_1,e_1\}$

In the Table 1, the items in the dataset are I={a1,b1,c1,d1,e1}and the transactions are T={T1,T2,T3,T4 and T5}. The items purchased by each customer are represented in each transaction such that a, c, d are the items purchased by the first customer as given in T1. The cardinality of the itemset is the number of items purchased by a single customer, whereas the cardinality of T1 is 3. The interesting patterns can be mined from the transactions given in the itemset. The minimum support of the itemset is identified by enumerating all the patterns extracted from the dataset. In the process of framing classification rules, the data items are represented as X and the class label is termed as C. The rules are defined using two measures including support and confidence. The support can be defined as the ratio of the A and B transactions to the transactions in the dataset (7).

$$\textbf{Degree of Support } (\mathbf{A} \rightarrow \mathbf{B}) = \frac{||\mathbf{A} \cup \mathbf{B}||}{|\mathbf{T}|} \quad (1)$$

The ratio of transactions that has A and B to the transactions that has A is given in equation (2)

$$\text{Degree of Support } (\mathrm{A} \rightarrow \mathrm{B}) = \frac{||\mathrm{A} \cup \mathrm{B}||}{|\mathrm{A}|} \quad (2)$$

The classification rules are generated similar to the association rules based on the traditional apriori algorithm such as $X \rightarrow C$, where X is the dataset and C is the target class. The filters are applied to sort the rules to make decisions for profitable business. In general, the apriori method includes all the candidate itemsets for rule generation, which become a tedious process in case of large dataset. In this paper, the Top N random forest classification is applied to forecast the high-utility itemset from Instacart market dataset. The search space is therefore pruned, which in turn reduces the consumption of execution time and space.

Remaining section of paper is ordered as follows. State of art techniques to excavate high-utility itemsets is surveyed in Section II. Proposed Top N random forest classification is described in Section III in a detailed manner. The results and the dataset utilized for experimentation in discussed in Section IV. The paper is concluded with a crisp summary in Section V.

2. Related Art

This section reviews the recent researches on high-utility itemset mining along with pros and cons. An efficient algorithm for high-utility itemset mining is introduced to reduce memory consumption and runtime of the transactional datasets. The unnecessary search space in the dataset is eliminated by revisiting the utility sub-tree and local itemset in sub tree. The upper boundary of time limit is computed using the array based counting approach. The memory consumption of the suggested algorithm is found to be low on comparison with the traditional itemset mining algorithms (8). The closed high utility itemsets are mined using the CLS miner algorithm that uses the list structure to organize the utilities. It also utilizes the pruning techniques to reduce the computational complexity. The closed high-utility itemsets facilitate lossless and compact illustration of frequently used itemsets in a list (9).

The recommendation of advertisements in the online websites has been suggested by the Particle Swarm Optimization (PSO) algorithm for frequent mining of itemsets. In PSO, current particles that are optimum are followed by the other populations and positions are changed frequently to determine the best optima position. The best value is chosen based on the local best. When a new position is reached by the particle, that position is updated for each time frame. The rule mining approach is considered as an effective method to map the relationship among the items in the dataset. A threshold based strategy is adopted in PSO algorithm which chooses the items which are higher than the threshold value (10). The itemsets from the real time transactional databases are extracted based on the maximum usage of each itemset using the maximal itemset mining algorithm. The search space of the dataset is reduced to improve the memory efficiency using the pruning techniques (11). The statistical learning techniques have been adopted to mine the high utility datasets using tree explainer method. The non-monotonic rules are applied for the fast mining of itemsets based on Artificial Intelligence (AI) techniques. The tedious algebraic problems are easily solved using the statistical approaches which in turn reduce the computational complexity (12).

The elimination of candidate generation helps in the effective mining of frequently used itemsets. A high utility miner without candidate generation is proposed to organize the utility lists in a quick manner. The lists of utilities are arranged in a horizontal manner to improve pruning mechanism (13). The frequent and high utility itemsets are identified using metaheuristic approaches including sward intelligence, ant colony optimization and genetic algorithms. These strategies assist in solving the combinatorial issues to locate the near optimal solutions, thereby finding itemsets in large spaces. It favors in diminishing the time complexity by pruning the large search spaces (14). An improved algorithm for efficient mining of high-utility itemset is introduced to fasten the extraction thereby decreasing the transaction count. A P-set structure is suggested to eliminate frequent scans in the transactions. The significant features are essential to make the mining algorithms suitable for real life applications such as retail store marketing (15).

The classes of regularities are considered to mine the itemsets in spatio temporal datasets. A threshold that specifies the minimum distance between two itemsets is predefined. The process of satisfying the anti-monotonic property is a critical task in spatio temporal data mining. All the itemsets in the dataset is evaluated using the first single scan algorithm (16). The behavior pattern of the customers can be identified using the transactions that take place in a database. The time and memory consumption are the major drawbacks in utility list based itemset mining. A buffer for the list is maintained to construct the segments in which the highly accessible itemsets are placed. The buffer incorporated list structure facilitates fast discovery of high utility itemsets (17). A survey has been conducted on a set of 10 high utility itemset mining algorithms in which d2HUP performed well in terms of runtime. The Efficient Frequent Itemset Mining (EFIM) algorithm is found to be highly memory efficient on comparison to the conventional approaches (18).

A novel algorithm namely, SKYMINE is designed to utilize non-dominated patterns for discovering the frequently accessed itemsets. The SKYMINE adopts breadth-first-search and depth-first-search to identify appropriate itemsets. The UP tree structure is replaced by the utility list structure for efficient mining. It reduces the memory consumption, runtime and search space size (19). Length and itemset utilities are considered to mine the average utility itemsets. The upper bound utilities, namely, average, tighter and looser bounds are computed to minimize the search area of itemsets. It reduces the times utilized for searching itemsets in the unnecessary areas. In addition, the researchers tend to worn on reducing the candidate pattern size (20). The various optimization techniques used in classification is proposed in (21-23). The low frequency itemsets play a vital role in assessing the profit of a business, using which the low frequency items can be recommended for offer sale. The high and low frequency itemsets are combined to generate accurate association rules for mining the profitable patterns (24). The frequent itemsets in a local region of the dataset greatly helps in attaining the global accuracy of mining. The items that are sold for a particular time span from a market is represented as local utility patterns. The items that change in a frequent manner are eliminated to achieve the accuracy in prediction (25).

3. Proposed System

This section describes in detail about the proposed Top N random forest model utilized to estimate probabilities of high-utility itemsets of purchased products. The itemsets with high probability scores are recommended using Top N variants. The overall description of the proposed model is given in Figure. 1.

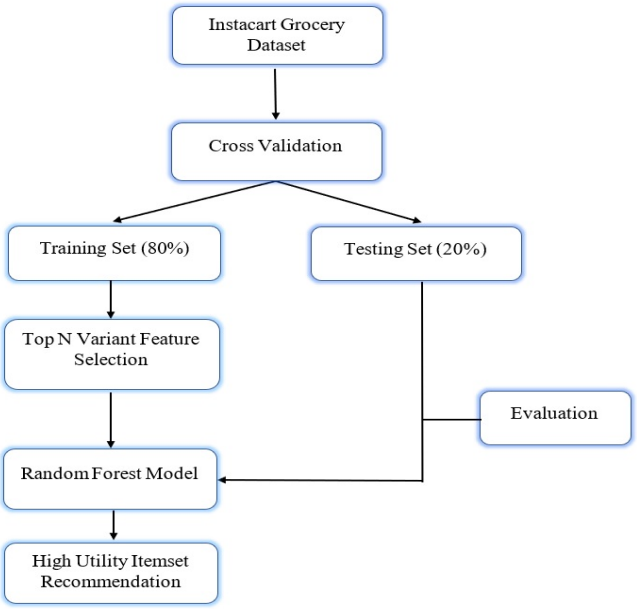

Figure 1. Overall flow of Top N variant Random Forest model.

A group of randomized subtrees are collected in a tree structure to generate a random forest model. An ensemble of decision-trees is generated to construct each tree structure in the random forest. The bootstrap sampling in performed on the training set through which ‘n’ random records are selected and sampled. Each node has a coordinate x which is selected according to the probability value of the ith feature. The features from the parent node are inherited by the child nodes. The trees in the random forest produce a response in accordance with set of values given to each tree. The random forest easily handles the missing values in the dataset.

(i) Bootstrap samples are derived from the training dataset.
(ii) Each sample is constructed as tree in which ‘n’ random features are selected to split a node into two children.

(iii) The tree is subdivided until the node size becomes minimum value.
(iv) An Out-Of-Bag (OOB) error rate us computed using the data that is not present in the bootstrap sample

In the proposed method, the features selected via the Top N variants are provided as input to the RF to train the classifier. The selection of appropriate features improves the accuracy of the model utilized for training. The training cases are split using bootstrap sampling in which the next split is selected according to the Gini index. The trees are fully grown until there is no decrease in the error. The main advantage of RF is that it can handle heterogeneous types of data and ease identification of outliers. It is not highly sensitive at the same time it has a large computational scalability.

3.1. Top N Variants Selection creation

The threshold of top N variants is selected from the highest probability of user – product pairs. The reparameterization of the threshold is represented as follows:

$$N_{Threshold} = \{(u,p)|P((u,p)) > P_0\} \quad (3)$$

The top variants facilitate the recommendation of high utility itemsets that are likely to be reordered again, still the variations results in large number of recommendation. Hence, the value of threshold probability to be set as 0.5 as given below.

$$N_{0.5} = \{(u,p)|P((u,p)) > 0.5\} \quad (8)$$

The skew predictions plays a vital role in recommending exact itemsets, in which skew is the ratio of negative classes in the training set to the positive classes in the training set. The skew of user-product pair is represented as:

$$N_{skew(s)} = n_s * P(y_+) = \frac{n_s}{1+skew} \quad (9)$$

Where,

$$P(y_+) = \frac{y_+}{y_+ + y_-} \quad (10)$$

In which, y_+ denotes the number of positive classes and y_- is negative classes in training dataset. The basket size and the mean of reordered items from a single basket by a user is computed as follows:

$$N_{basket} = \sum_{u \epsilon U} \overline{b(u)} \quad (11)$$

Where $\overline{b(u)}$ is the mean of basket size.

$$N_{reordred} = \sum_{u \epsilon U} \overline{r(u)} \quad (12)$$

In which $\overline{r(u)}$ denotes the mean of reordered items.

4. Results and Analysis

The dataset consists of groceries purchased by the customers in the Instacart online application. The purchase orders of customers are given in each transaction with a sample of 200,000 Instacart users. The sequence of purchase is provided with around 100 orders for each customer. The dataset contains the week, hour and day in which the order is placed. Each product has a unique product id as well as department id along with the time and day of purchase. The data analysis is done using python using pandas, seaborn and matplotlib packages. The datasets are processed using tensorflow framework due to its huge volume. The ultimate goal is to predict the next set of items to be purchased by the customer. Hence, machine learning algorithm, namely, Top N random forest classifier is suggested for next order prediction, thereby increasing the profit of the market. The glimpse of the transaction table along with the features is shown in Table 2.

The dataset is analyzed for the presence of prior orders that are repeated again by the customers. It is found that at least 3 products which have been ordered already are ordered again. The plot for prior orders ordered by the customers is depicted in Figure 2(a). The proportion of reordered items is shown in Figure 2(b).

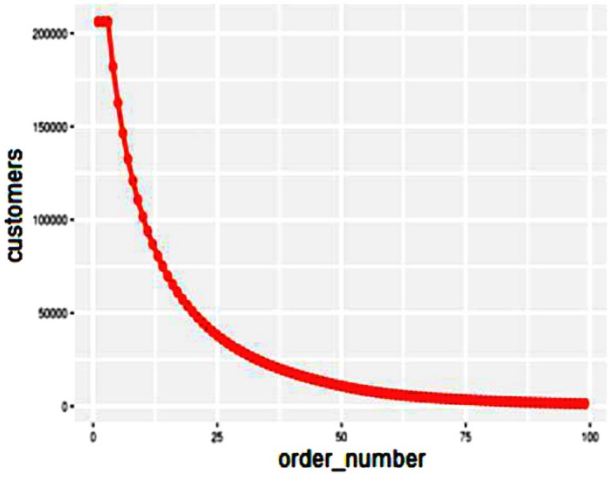

Figure 2(a). Occurrence of prior items in further purchase orders

It is evident from the ceiling effect that the item that is most frequently ordered are purchased reordered. The dataset is scanned and sorted in the descending order to identify the frequency. The transactions and the items are scanned during second iteration which forms a transaction tree by sorting frequent items in the ascending order. The total items and reordering items probability association relationship is shown in Figure 3(a). The reordering probability of unique products in the dataset is given in Figure 3(b).

Table 2. Sample transactions in the dataset

variables 8		
$ORDER_ID	<INT>	263878,2398795,4567823, 2464736,4413…
$USER_ID	<INT>	1,3,1,1,2,1,4,1,2,1,3,2,2,…
$EVAL_SET	<CHR >	"prior","prior","prior","prior"…
$ORDER_NO	<INT>	1,5,3,7,5,6,5,8,11,10,13, 1,2,…
$ORDER_DOW	<INT>	2,4,4,4,5,2,1,1,2,4,5,5,5,…
$ORDER_HOUR	<INT>	8,8,12,8,16,8,10,13,15,9, 8,10,…
$DAYS_SINCE_ PRIOR_ORDER	<DBL >	NA,16,22,28,28,29,10,14, 0,20,…

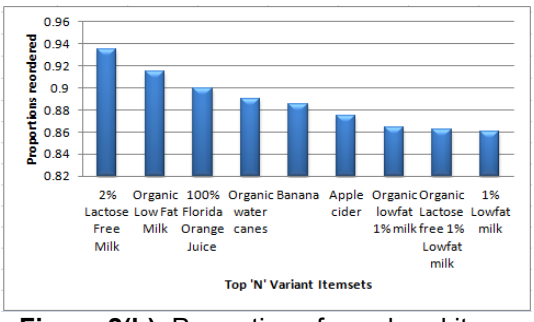

Figure 2(b). Proportion of reordered items

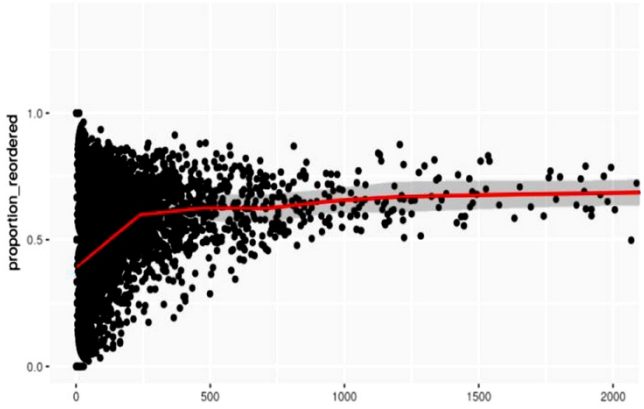

Figure 3(a). Probability reordering items

4.1. Feature Selection

Top N-variants are selected to recommend high-utility itemsets according to probability of purchase. The importance of the feature is computed in accordance with equation (3) that is the skew between the pair of user and product. The measure feature importance greatly helps in improving the performance of the classifier that maps the interactions between the user and feature of products.

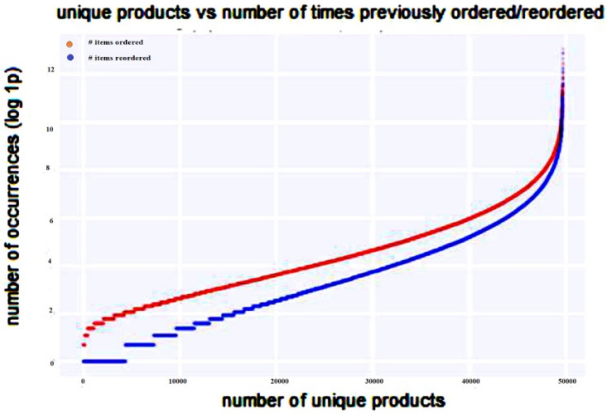

Figure 3(b). Probability reordering unique products

4.2. Receiver Operating Characteristic curve

Despite of predicting the products to which class it belongs to, it is better to predict the probabilities of high utility itemsets for reducing the complexity. The false negatives and false positives are compared to interpret the probabilities of the observations using various threshold values. The ROC curve and the precision recall curve is plotted to measure the probability forecast in classification problems. The trade-off between the actual and predictive values of true positives and false positives are represented in the ROC curve plotted in Figure 4. The dataset is segregated as 80% for training and 20% for testing. It is seen from Figure 4 that the training set attained an accuracy of 86% and the testing set resulted in 83% accuracy. The slight difference in accuracy shows that the model is good enough for prediction.

4.3. Precision and Recall

The two most important measures to determine the classification model performance is precision and recall. Precision, a positive predictive representation can be defined as number of true positives to total of true and false positives. The value of recall is equivalent to the sensitivity of the model which is defined as the ratio of number of true positives to aggregation of false negatives and true positives. Significance of precision and recall is that it eliminates the true negative, thereby considering only the positive predictions. The below Figure 5 represent the precision and recall measures of the model. The accuracy obtained from the RF model is compared with SVM and the results obtained are shown in the below Figure 6. From Figure 6 it is understood that the existing Support Vector Model gains an accuracy of 82% for the top ‘N’ variant itemsets that is against 87% for the proposed one. The cost benefits of

utilizing RF model gains advantages in market basket analysis for recommending the items that are reordered usually.

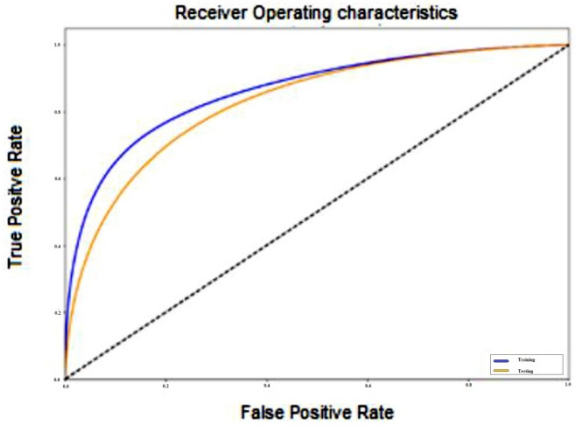

Figure 4. Receiver Operating Characteristic Curve

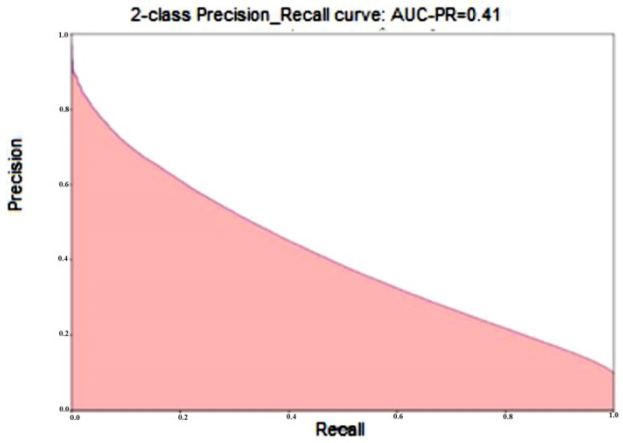

Figure 5. Precision and Recall Curve

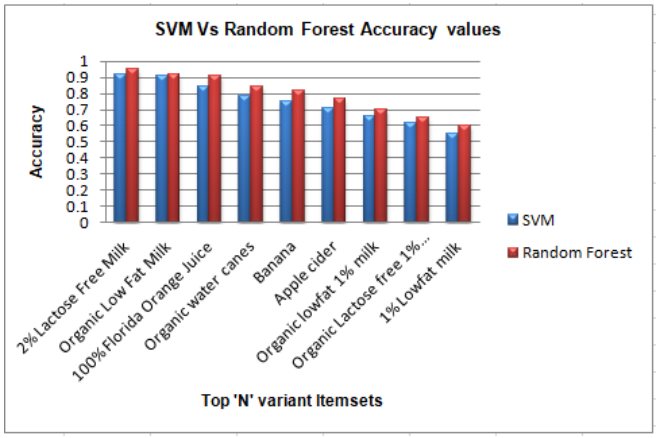

Figure 6. Proposed RF model vs SVM in terms of accuracy

5. Conclusion and Future Work

Top N variants Random Forest model is proposed to predict the high-utility itemsets in the dataset. The Instacart market dataset that has more 200,000 transaction records are utilized to evaluate the model. The huge dataset is handled using tensor flow framework. The entire dataset is analyzed to identify the appropriate features from the dataset for training the model. The probability of each feature is computed according to which the features are selected. Top N variants are selected and they are given as input to the random forest model to forecast the high utility itemsets. The model attained the testing accuracy of 83% in such a way that it could suggest the high itemsets in an efficient way. In future, the measures will be taken to perk up the accuracy further.

References

[1] Anshari, Muhammad, et al. Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics 15.2 (2019): 94-101.

[2] Buenaño-Fernandez, Diego, William Villegas-CH, and Sergio Luján-Mora. The use of tools of data mining to decision making in engineering education—A systematic mapping study. Computer Applications in Engineering Education 27.3 (2019): 744-758.

[3] Balakrishna, E., B. Rama, and A. Nagaraju. Efficient Mining of Negative Association Rules Using Frequent Item Set Mining. First International Conference on Artificial Intelligence and Cognitive Computing. Springer, Singapore, 2019.

[4] Khan, Mohiuddin Ali, Sateesh Kumar Pradhan, and Huda Fatima. An Efficient Technique for Apriori Algorithm in Medical Data Mining. Innovations in Computer Science and Engineering. Springer, Singapore, 2019. 187-195.

[5] Hossain, Maliha, AHM Sarowar Sattar, and Mahit Kumar Paul. Market Basket Analysis Using Apriori and FP Growth Algorithm. 2019 22nd International Conference on Computer and Information Technology *(ICCIT)*. IEEE, 2019.

[6] Nguyen, Giang, et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review 52.1 (2019): 77-124.

[7] Buaton, Relita, et al. Decision Tree Optimization in Data Mining with Support and Confidence. Journal of Physics: Conference Series. Vol. 1255. No. 1. IOP Publishing, 2019.

[8] Zida, Souleymane, et al. EFIM: a fast and memory efficient algorithm for high-utility itemset mining. Knowledge and Information Systems 51.2 (2017): 595-625.

[9] Dam, Thu-Lan, et al. CLS-Miner: efficient and effective closed high-utility itemset mining. Frontiers of Computer Science 13.2 (2019): 357-381.

[10] Keerthi, M., et al. Mining High Utility Itemset for Online Ad Placement Using Particle Swarm Optimization Algorithm. International Conference On Computational Vision and Bio Inspired Computing. Springer, Cham, 2019.

[11] Nguyen, Trinh DD, Quoc-Bao Vu, and Loan TT Nguyen. Efficient algorithms for mining maximal high-utility itemsets.

2019 6th NAFOSTED Conference on Information and Computer Science (NICS). IEEE, 2019.
[12] Shakerin, Farhad, and Gopal Gupta. Induction of Non-Monotonic Rules From Statistical Learning Models Using High-Utility Itemset Mining. arXiv preprint arXiv:1905.11226 (2019).
[13] Qu, Jun-Feng, Mengchi Liu, and Philippe Fournier-Viger. Efficient algorithms for high utility itemset mining without candidate generation. High-Utility Pattern Mining. Springer, Cham, 2019. 131-160.
[14] Djenouri, Youcef, et al. Metaheuristics for Frequent and High-Utility Itemset Mining. High-Utility Pattern Mining. Springer, Cham, 2019. 261-278.
[15] Nguyen, Loan TT, et al. Mining high-utility itemsets in dynamic profit databases. Knowledge-Based Systems 175 (2019): 130-144.
[16] Kiran, R. Uday, et al. Discovering spatial high utility itemsets in spatiotemporal databases. Proceedings of the 31st International Conference on Scientific and Statistical Database Management. 2019.
[17] Duong, Quang-Huy, et al. Efficient high utility itemset mining using buffered utility-lists. Applied Intelligence 48.7 (2018): 1859-1877.
[18] Zhang, Chongsheng, et al. An empirical evaluation of high utility itemset mining algorithms. Expert Systems with applications 101 (2018): 91-115.
[19] Lin, Jerry Chun-Wei, et al. Mining of skyline patterns by considering both frequent and utility constraints. Engineering Applications of Artificial Intelligence 77 (2019): 229-238.
[20] Wu, Jimmy Ming-Tai, et al. TUB-HAUPM: Tighter upper bound for mining high average-utility patterns. IEEE Access 6 (2018): 18655-18669.
[21] Suganya, E., et al. Mobile Cancer Prophecy System to Assist Patients: Big Data Analysis and Design. Journal of Computational and Theoretical Nanoscience 16.8 (2019): 3623-3628.
[22] Sountharrajan, S., et al. Automatic classification on bio medical prognosis of invasive breast cancer. Asian Pacific Journal of Cancer Prevention*: APJCP* 18.9 (2017): 2541.
[23] Sountharrajan, S., et al. Dynamic Recognition of Phishing URLs Using Deep Learning Techniques. Advances in Cyber Security Analytics and Decision Systems. Springer, Cham, 2020. 27-56.
[24] Wu, Jimmy Ming-Tai et al. Mining Association rules for Low-Frequency itemsets. PloS one vol. 13,7 e0198066. 23 Jul. 2018, doi:10.1371/journal.pone.0198066
[25] Fournier-Viger, Philippe, et al. Mining local high utility itemsets. International Conference on Database and Expert Systems Applications. Springer, Cham, 2018