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/_____

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

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,…

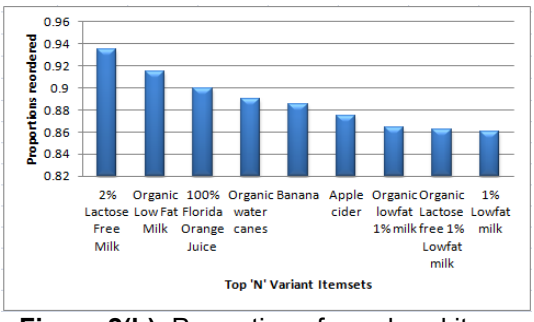

Figure 2(b). Proportion of reordered items

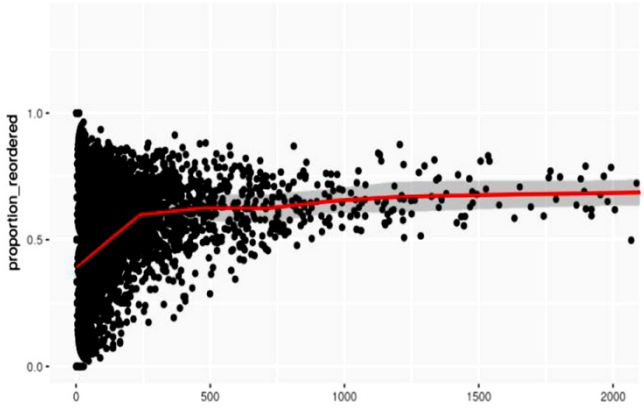

Figure 3(a). Probability reordering items

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.

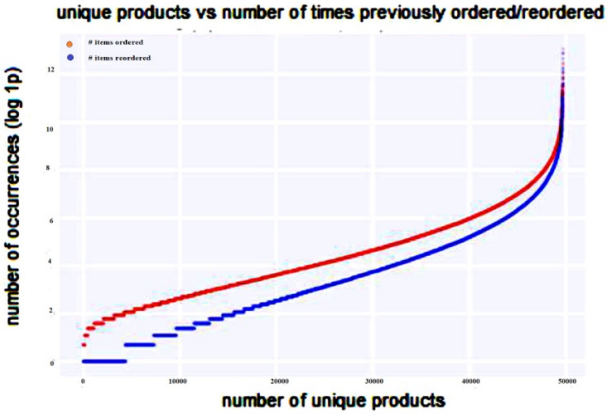

Figure 3(b). Probability reordering unique products

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.

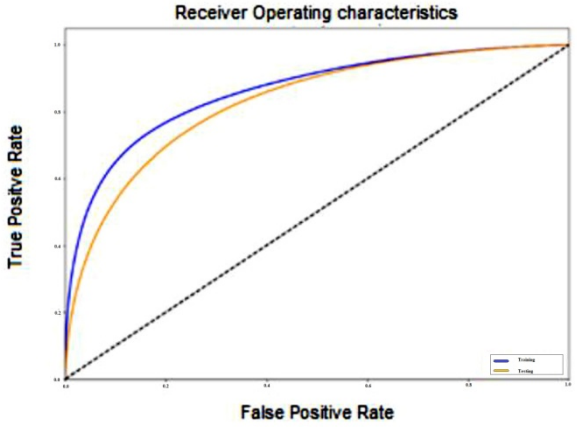

Figure 4. Receiver Operating Characteristic Curve

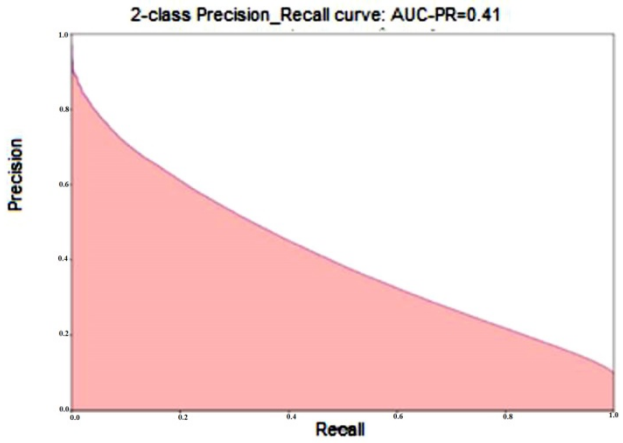

Figure 5. Precision and Recall Curve

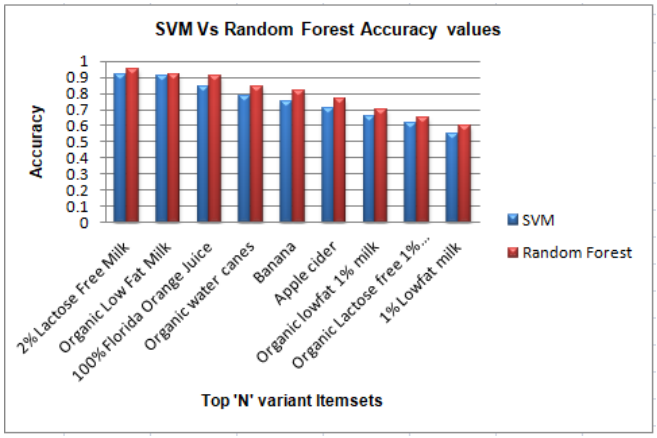

Figure 6. Proposed RF model vs SVM in terms of accuracy

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