International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072

Analyzing BPFP for Pattern Mining

Akansha Pandey[1], Shri Prakash Dwivedi[2], H. L. Mandoria[3]

[1]Student, Dept. of Information Technology, College of Technology, GBPUAT, Uttarakhand, India
[2]Asst. Professor, Dept. of Information Technology, College of Technology, GBPUAT, Uttarakhand, India
[3]Professor, Dept. of Information Technology, College of Technology, GBPUAT, Uttarakhand, India

---

Abstract - *A Data mining consists of various techniques which can be used to find patterns from a huge database. Because of the continuous collection of massive amount of data many industries have become interested in frequent pattern mining techniques. The discovery of interesting relationships among a large amount of records can help in decision making process. This paper focuses on analyzing the performance of frequent pattern mining i.e. BPFP algorithm by varying the trained dataset and the maximum number of devices in terms of root mean square error. Also the association rules are measured for different transactions in terms of lift.*

Key Words: **Frequent Pattern mining, BPFP, Lift, Root mean square error.**

1. INTRODUCTION

Pattern is a regular and intelligible form or sequence discernible in the way in which something happens or is done. Frequent patterns are the patterns that appear frequently in a database. The idea of frequent pattern mining is to search for knowledge, which means consistencies, frequency, rules and structures hidden in the data. This task is a subfield of information technology called knowledge discovery or sometimes data mining. This knowledge will help in making decisions and conclusions that lead to value creation for both the user and the owner of the data. For instance, the purchase information collected by a supermarket chain may help the supermarket to adjust product offering and availability to better suit the needs of their customers. Moreover, bus and train companies can use recorded passenger data to help plan bus services to run more often where needed.

Finding frequent patterns helps in mining correlations, clustering, data classification, associations and many other interesting relationships among data.

2. Balanced Parallel Frequent Pattern Algorithm

The BPFP (Balanced Parallel Frequent Pattern Algorithm) approach involves the following steps:

1. The whole database is first divided into successive parts and stored on different devices.

2. The next step is to create the F-list which contains the list of frequent items sorted in descending order according to the frequency on different devices.

3. After getting the list of frequency of the itemset the next step is to generate association rules and create FP Growth tree.

3. Motivation

As the accuracy of frequent pattern matching is a very difficult task in the data mining. To overcome these problems many researchers have developed the efficient algorithms that aim to reduce the error and standard deviation. In the previous work, the BPFP (Balanced Parallel FP) algorithm is not widely used and the mean square errors and lift parameters were not considered for analysis. Hence, in this research paper the algorithm is analyzed on the basis of these parameters.

4. Working of the method

The method applied follows the following steps:

1. First step is to divide the whole database into two parts namely trained data set and test data set. The trained data set consists of sample of the whole database. And test data set consists of the remaining data which is used to check the correctness of the algorithm.
2. Then for each pair of input variables of the training set is fed to the devices and the algorithm works on that set according to the number of layers set. The output is set as observation. Then the root mean square error and standard deviation is computed for the output result.
3. After computing the result for trained dataset the algorithm is applied on test data set

and the root mean square error and standard deviation are computed for the same and compared with the trained dataset.

4. For evaluating the association rule, the minimum support threshold and minimum confidence threshold are set. Then for the input dataset frequent pattern algorithm is applied.
5. The lift, support and confidence are calculated for each rule to get to know the interestingness of the rules.

5. Results and Discussion

This section discusses the results of the algorithm and analyzes them. The coding has been done in Matlab.

The Performance parameters taken for analyses are as follows:

i. Lift

Lift in data mining is defined as a measure for evaluating the performance of the targeting model or an association rule as having an emphasized response with respect to the whole population, measured against a random choice targeting model. With the computation of lift it can be said that targeting model which is selected is good if the response within the target is much better than the average.

ii. Root Mean Square Error

Root-mean-square error (RMSE) is a measure that represents the differences between values predicted by a targeted model and the values actually observed..

5.1 Analyzing Balanced Parallel Frequent Pattern Algorithm based on varying number of maximum devices for train data and test data

This part gives the analysis of results on the algorithm discussed. In this study the performance of frequent pattern mining method i.e. BPFP algorithm on different data sets and different thresholds is discussed and analyzed.

5.1.1. Analyzing average root mean square for different number of devices and varying train data

i. Train Data

In **Fig 1:** 0.75, 0.85, 0.90 represent the train data taken. The graph shows that only the selection of trained data set is not enough for the accuracy of the result. The number of devices also plays a great role in the results. More the number of devices lesser is the root mean square error for the trained data.

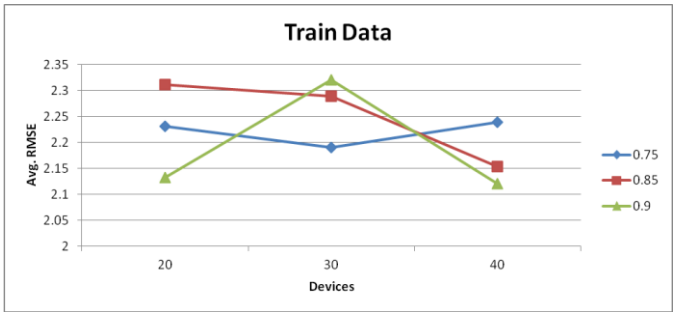

Chart 1: Root mean square error for varying train data an different number of devices

ii. Test Data

Fig 2: shows that the overall root mean square error decreases with increase in the number of devices. It can also be seen that the test data has less error as compared to the train data.

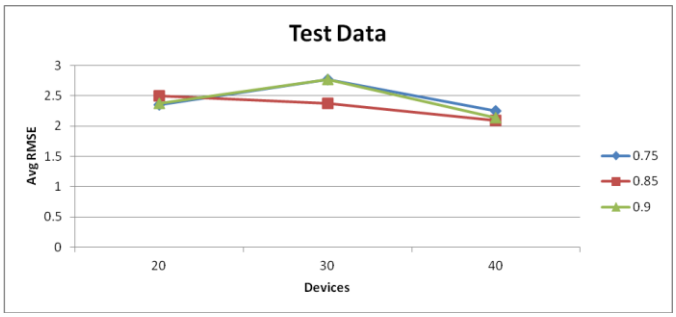

Chart 2: Root mean square error for test data at different values of train data and different number of devices

5.2. Analyzing the interestingness of association rules in terms of lift

Table 1: Transactions - Itemset

Transactions	Itemset
T1	[1,2,5]
T2	[2,4]
T3	[2,3]
T4	[1,2,4]
T5	[1,3]
T6	[2,3]
T7	[1,3]
T8	[1,2,3,5]
T9	[1,2,3]

Table 2: Extracting rules at minimum support threshold = .20, minimum confidence threshold = 0.07

Rules	Support	Confidence	Lift
Rule #1: 2 --> 5	0.22222	0.28571	1.2857
Rule #2: 5 --> 2	0.22222	1	1.2857
Rule #3: 1 --> 5	0.22222	0.33333	1.5
Rule #4: 5 --> 1	0.22222	1	1.5
Rule #5: 2 --> [1, 5]	0.22222	0.28571	1.2857
Rule #6: 1 --> [2 5]	0.22222	0.33333	1.5
Rule #7: [2 1] --> 5	0.22222	0.5	2.25
Rule #8: 5 --> [2 1]	0.22222	1	2.25
Rule #9: [2 5] --> 1	0.22222	1	1.5
Rule #10: [1 5] --> 2	0.22222	1	1.2857
Rule #11: 2 --> 4	0.22222	0.28571	1.2857
Rule #12: 4 --> 2	0.22222	0.28571	1.2857
Rule #13: 1 --> 3	0.44444	0.66667	1
Rule #14: 3 --> 1	0.44444	0.66667	1

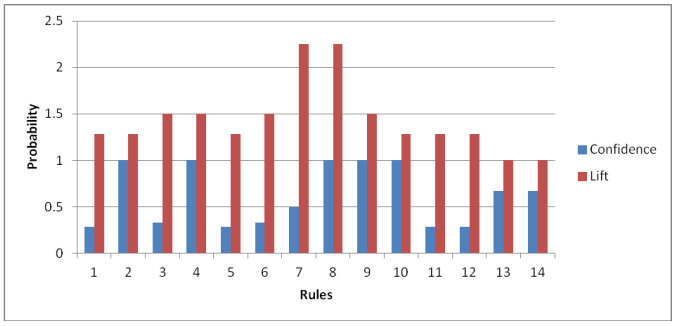

Chart 3: Rules vs. Probability

The above Chart 3 shows that the minimum support threshold and minimum confidence threshold are not enough for checking the interestingness of a pattern. Another parameter that is lift must also be taken into consideration as some of the rules were satisfying the confidence and support threshold criteria but the lift for them was 1 which represents the independency of the itemsets of the rules.

6. CONCLUSIONS

The performance of pattern mining i.e. BPFP was analyzed by varying the trained dataset and number of layers in terms of RMSE and SD. It can be concluded that by increasing the number of devices the difference of RMSE and SD between the trained dataset and target dataset decreases.
In future this algorithm can be implemented for different data mining tasks such as clustering, classification, etc. Also the precision and other factors can also be taken into consideration for improvement.

REFERENCES

[1] Agrawal, R., & Srikant, R. 1994. Fast algorithms for mining association rules. Proceedings of 1994 International Conference Very Large Data Bases, 487–499

[2] Bhat, Ramya S., & Beham, A. Rafega 2016. Comparative Study on Algorithms of Frequent Itemset Mining. International Journal of Computer Science and Mobile Computing, 271-275.

[3] Bose, Subrata, & Datta, Subrata 2015. Frequent Pattern Generation in Association Rule Mining using Weighted Support. 3rd International Conference on Computer, Communication, Control and Information Technology, 1-5.

[4] Buehrer, Gregory, Jr., Roberto L. de Oliviera, Fuhrey, David, & Parthasarathy Srinivasan 2015. Towards Parameter-Free and Parallel Itemset Mining Algorithm in Linearithmic Time. ICDE Conference 2015, 1071-1082.

[5] Burdick, Doug, Calimlim, Manuel, & Gehrke, Johannes 2001. MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases. IEEE Transactions on Knowledge and Data Engineering, 1490-1504.

[6] Chung, Soon M., & Luo, Congnan 2015. Parallel Mining of Maximal Frequent Itemsets from Databases. Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 134-139.

[7] Grahne, Ghosta, & Zhu, Jianfei 2005. Fast Algorithms for Frequent Itemset Mining Using FP-Trees. IEEE Transactions on Knowledge and Data Engineering, 1347-1362.

[8] Han, Jiawei, Pei, Jian, Yin, Yiwen, & Mao Runying 2004. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 53-87.

[9] Han, Jiawei, Cheng, Hong, Xin, Dong, & Yen, Xifeng 2007. Frequent pattern mining: current status and future directions. Data Mining Knowledge Discovery, 55-86.

[10] Li, Haoyuan, Wang, Yi, Zhang, Dong, Zhang, Ming, & Y. Chang, Edward 2008. PFP: Parallel fp-growth for

query recommendation. Proceedings of the 2008 ACM conference on Recommender systems, 107-114.

[11] Liu, Yanxi 2010. Study on Application of Apriori Algorithm in Data Mining. 2010 Second International Conference on Computer Modeling and Simulation, 111-114.

[12] Qiu, Yong, Lan, Yongjie, & Xie, Qing-Song 2004. An Improved Algorithm of Mining from Fp-Tree. Proceedings of the Third International Conference on Machine Learning and Cybernetics, 1665-1670.

[13] R. Zaiane, Osmar, El-Hajj, Mohammad, & Lu, Paul 2001. Fast Parallel Association Rule Mining without Candidacy Generation. Proceedings of the 2001 IEEE International Conference on Data Mining, 665-668.

[14] Rao, G. Nageshwar, & Gurram, Suman Kumar 2011. Mining frequent item sets without candidate generation using FP-Trees. International Journal of Computer Science and Information Technologies, 2677-2685.

[15] S. Bhat, Ramya, Rafega Beham, A. 2016. Comparative Study on Algorithms of Frequent Itemset Mining. International Journal of Computer Science and Mobile Computing, 271-275.

[16] S., Patel Tushar 2015. Performance Analysis of Frequent Itemset Finding Techniques using Sparse Datasets. IEEE International Conference on Computer, Communication and Control (IC4-2015), 1-4.

[17] Savasere, A., Omiecinski, E., & Navathe, S. 1995. An Efficient Algorithm for mining Association Rules in large databases. Proceedings of International Conference on Very Large Data Bases, 432–443.

[18] Shah, Arpan H., & Patel, Pratik A. 2015. Optimum Frequent Pattern Approach for Efficient Incremental Mining on Large Databases using Map Reduce. International Journal of Computer Applications, 25-29.

[19] Singh, Sanasam Ranbir, Kr. Patra, Bidyut, & Giri, Debasis. Mining Frequent Closed Itemsets using Conditional Frequent Pattern Tree. IEEE India Annual Conference 2004. Indlcon 2004, 501-504.

[20] Su, Ja- Hwung, & Lin, Wen Yang 2004. CBW: An Efficient Algorithm for Frequent Itemset Mining. Proceedings of the 37th Hawaii International Conference on System Sciences – 2004, 1-9.

[21] Tseng, Fan-Chen, Hsu, Ching-Chi, & Chen, Henry 2001. Mining Frequent Closed Itemsets with the Frequent Pattern List. Proceedings of the 2001 IEEE International Conference on Data Mining, 653-654.

[22] Wang, Lei, Fan, Xing-Juan, Liu Xing-Long, & Zhao, Huan 2012. Mining Data Association Based On A Revised Fp-Growth Algorithm. Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, 91-95.

[23] Wen, Lie 2004. An Efficient Algorithm for Mining Frequent Closed Itemset. Proceedings of the 5th World Congress on Intelligent Control and Automation, 4296-4299.

[24] Yun, Unil, & J. Leggett, John 2016. WFIM: Weighted Frequent Itemset Mining with a weight range and a minimum weight. Proceedings of 2005 SIAM International Conference on Data Mining, 636-640.