Regular Pattern Mining on Crime Data Set using Vertical Data Format

D. Ravikiran
Research Scholar
Acharya Nagarjuna University
Nagarjuna Nagar, Guntur

S. V. N. Srinivasu, PhD
Research Supervisor
Acharya Nagarjuna University
Nagarjuna Nagar, Guntur

ABSTRACT
At Present Mining regular patterns in the data are an emerging research area. The huge escalations in the records and scopes of available database regular pattern mining are interesting problem in present days. Regular pattern mining is important criterion for measuring several applications like crime analysis, market basket analysis and web analysis .To mine regular patterns we use vertical data format technique in crime database approach achieves better performance. Not only frequency, the regularity of the item also can be consider as emerging factor in data mining research. Here we proposed a model to mine regular sort of crimes in crime database using vertical data format.

Keywords
Regular pattern mining, frequent pattern mining, Crime data set, Vertical data format, Association rule

1. INTRODUCTION
Crime is one of the problem facing by the people, Wrongdoing is one of the worldwide difficulties and numerous wrong doing recognition intrigues in Law Requirement offices where web information constantly offer profitable and fitting data for law organization. So the information in designed upon the time and information premise of wrongdoing and additionally in light of rate of the wrongdoing events i.e. no of violations happened in their individual circumstances or the occasions or between some time compass. Along these lines much of the time done violations are to be seen to make the data over law organization to make open wellbeing what's more, in security.

Along these lines incessant example mining is utilized to design the data over wrongdoing database. The incessant example mining gives the data of regularly events which may not generally speak to the noteworthiness of an example. The fundamental point in successive example mining is it doesn't make the state of event of wrong doing. It also a challenge for the security agencies for public safety, these security agencies tackle the problem by using many techniques like rp tree and vertical data format. Hence by applying some mining techniques helps them to predict future and go accordingly and will get benefits by knowing the future.

Generally data generated is continuous and unbounded and managing these data and identifying the associations is difficult task. Mining according to the user requirement is challenging task and therefore a pattern is called frequent if its occurrence frequency in the database exceeds the user-given support threshold. In case of frequent pattern mining on crime database is to identify the frequent crimes in that area with the algorithms like Apriori algorithm and FP-growth algorithm but while this frequent pattern mining techniques a large amounts of candidate keys will be generated and also it fails to cover the regular patterns because it focus only on the higher frequency patterns. To find regularity we can use vertical data format and rp tree. In rp tree it requires two database scan where as in vertical data format it uses only one database scan.

Tanbeer et al. [1] have proposed a tree based data-structure, called RPS-tree that captures user- given regularity threshold and mines regular patterns in a data stream with the help of FP-growth algorithm and conditional pattern bases and corresponding conditional trees. First, they constructed RPS-tree consists of one root node referred to as “null” and a set of item-prefix sub- trees called children of the root. Each node in an RPS-tree represents an item set in the path from the root up to that node. The RPS-tree maintains the occurrence information of all transactions in the current window with the tree structure. Also RPS-tree maintains two types of nodes called ordinary nodes and tail nodes. Nodes of both types explicitly maintain parent, children and node traversal pointers. In addition each tail node maintains a tid-list and a tail-node pointer. The tail- node pointer points to either the next tail node in the tree if any, or “null”. Then they construct an item header table called RPS-table consists of each distinct item in the current window with relative regularity and a pointer pointing to the first node in the RPS-tree that carries the item. RPS-table of a RPS-tree consists of three fields, they are item name (i), regularity of i (r), and a pointer to the RPS-tree for i(p). Similar to FP-growth mining, they mine the RPS-tree of decreasing size to generate regular patterns by creating conditional pattern-bases and corresponding conditional trees. Serial crime detection with the consideration of class imbalance problem is a novel approach[2] Periodic patterns [3], [4] and parallel patterns are also closely related with Regular patterns. Periodic pattern mining in time-series data focuses on the cyclic behavior of patterns either in whole or some part of time-series. Although periodic pattern mining is closely related with our work, it cannot be applied directly to mine regular patterns from a data stream because it process with either time-series or sequential data.In data mining, one of the most important techniques is Association rule mining. It was first introduced by[5].Cyclic patterns[6] are similar to the regular patterns.It extracts frequent patterns, correlations, associations among sets of items in databases. The main drawback with the classical Apriori algorithm is that it needs repeated scans to generate candidate set. After that Frequent pattern tree[7] and FP-growth[8] algorithm.

2. MINING REGULAR PATTERNS ON CRIME DATASET
Let C= {c1, c2, c3, c4…… cn} be a set of items. A set {c1,

c2, c3….. cn}⊆ C is called an item set (or pattern).

Period of x in transaction:Let t^i_x and t^x_{i+1} are two transactions ids, the number of transaction between t^x_i and t^x_{i+1} is defined as period of x, say P^x where $P^x = t^x_{i+1} - t^x_i$.

Consider first transaction as t_f as a null transaction i.e t_f=0 and the last transaction is t_l.

Regularity of x in transaction

Let $P^{x\ be}$ set of all periods of 'X' in the Transaction i.e. $P^w(X)$= $\{p^x_1, p^x_2, p^x_3, \ldots\ldots\ldots\ldots p^x_m\}$ where 'm' is the higher transaction number for 'X' appears in particular transaction

We say the pattern is regular if it occurs in the specified period otherwise it is not regular pattern. Regularity of pattern depends on the occurrence behaviour of patterns in a specified transaction.

Regular Pattern

Let X= $\{x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_m\}$ be a set of regular items and λ is the minimum regularity threshold. If Reg(x) is less than minimum regularity threshold then x is said to be regular item.

In this section we describe mining regular patterns in crime database with vertical data format structure which requires only one database scan. In this regard we consider a crime dataset contains crime transactions. Let us consider the minimum regularity threshold value, δ=5.Firstly, we have to consider each item in the dataset, for each item we have to write in how many transactions that particular item is presented, after that consider the minimum and maximum value. The minimum value is zero and the maximum is number of transaction in the table. In this transaction Tmin =0 and Tmax =8.

RCP Algorithm

Input: C D, λ

Output: Set of regular items

Procedure

1. Consider Crime Dataset CD.
2. Convert CD into vertical dataformat.
3. Let $X_i \in CD$, $X_i \subseteq$ k-itemset.
4. P X_i = 0 for all Xi
5. For every Xi calculate periodicity
6. P X_i = P X_i +1 - P X i
7. reg(X_i) = max(P X_i)
8. if reg(X_i) < = λ
9. X_i is regular itemset
10. Else
11. Delete X_i

In this algorithm consider a crime dataset and convert the crime dataset in to vertical data format. X_i be the set of items in the crime dataset, X_i be the subset of k itemset. Reg(X_i) is the maximum of P X_i , and λ is the minimum regularity threshold frequency if Reg(X_i) <= λ then X_i said to be regular item set and the remaining items are not regular Itemsets of CD with their periodicities in Table 1 and their regularities are available in Table 2. Ki, Mu, Ru, Bu, Ro, Ar, Sl are some crime datasets are consider in Table 1.For instance itemset (ki) is showed up in exchanges (1,2,3,4,6,7) and their periodicity values P^x (d) = { 1,1,1,1,2,1,1}.Regularity quality is 2 i.e., max(1,1,1,1,2,1,1) = 2. In our running case the base consistency limit is λ = 4. The itemsets which are having the consistency is not as much as or equivalent to least general limit are standard things. In this manner things {BU, KI, RO, MU} are customary itemsets and itemset {SL,AR}is not a standard itemset which are appeared in Table 2.

[Table 1]

ID	**TRANSACTION**
CT1	KI, RO
CT2	KI, MU, RO
CT3	KI, MU, RO, BU
CT4	KI, RO, BU
CT5	RO, MU, BU
CT6	KI, SL, AR
CT7	KI, SL, RO
CT8	RO, AR, SL

[Table 2]

Item set	**TID**	**PATTERN (tmax-tmin)**	**REG**
KI	1,2,3,4,6,7	(1,1,1,1,2,1,1)	2
RO	1,2,3,4,5,7,8	(1,1,1,1,1,2,1)	2
MU	2,3,5	(2,1,2,3)	3
BU	3,4,5	(3,1,1,3)	3
SL	6,7,8	(6,1,1)	6
AR	6,8	(6,2)	6

Subtract the minimum value and first transaction ID for each data item. The remaining values are the difference between the transactions. If the maximum value is not present in the transaction, the last transaction ID is subtracted from the maximum value. Now the value of the item is maximum number. Now, we assume the minimum threshold frequency value. If the transaction difference is less than are equal to the minimum threshold frequency value then those are said to be regular which is shown in Table 2. These items BU, KI, RO, MU are regular in single set. We can find two item data set and three item data set in the crime database. In two item dataset we have to consider two items in the dataset, for each two items we have to write in how many transactions that particular items are presented, after that considers the minimum and maximum value. The minimum value is zero and the maximum is number of transaction in the table. Subtract the minimum value and first transaction ID for each data item. The remaining values are the difference between the transactions. If the maximum value is not present in the transaction, the last transaction ID is subtracted from the maximum value. Now the value of the item is maximum

number. Now, we assume the minimum threshold frequency value. If the transaction difference is less than are equal to the minimum threshold frequency value then those are said to be regular then we will get two item dataset which is shown in Table 3. The regular items in double set are (KI, RO), (KI, BU), (RO, MU), (RO, BU), (MU, BU).

[Table 3]

Item set	TID	PATTERN (tmax-tmin)	REG
KI, RO	1,2,3,4	(1,1,1,1,4)	4
KI, MU	2,3	(2,1,5)	5
KI, BU	3,4	(3,1,4)	4
RO, MU	2,3,5	(2,1,2,3)	3
RO, BU	3,4,5	(3,1,1,3)	3
MU, BU	3,5	(3,2,3)	3

We can find three item data set in the crime database. In three item dataset we have to consider three items in the dataset which are regular, for each triple item we have to write in how many transactions that particular three items are presented, after that consider the minimum and maximum value. The minimum value is zero and the maximum is number of transaction in the table. Subtract the minimum value and first transaction ID for each data item . The remaining values are the difference between the transactions. If the maximum value is not present in the transaction, the last transaction ID is subtracted from the maximum value..Now the value of the item is maximum number. Now, we assume the minimum threshold frequency value. If the transaction difference is less than are equal to the minimum threshold frequency value then those are said to be regular then we will get two item dataset which is shown in Table 3. The regular items in triple set are (KI, RO, BU), (RO, MU, BU) in Table 4.

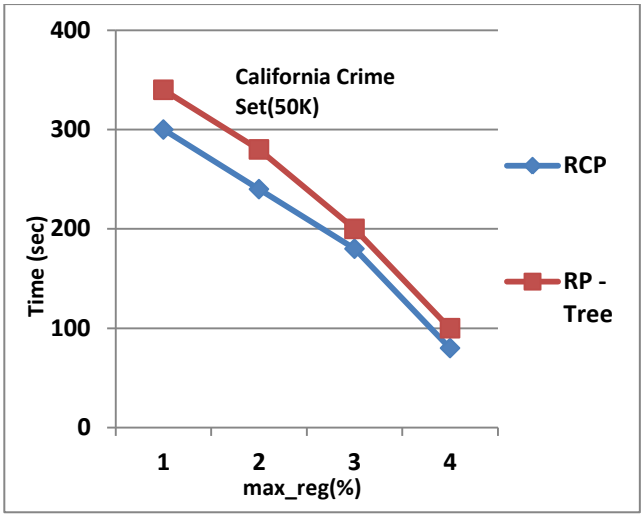
[Table 4]

Item set	TID	PATTERN (tmax-tmin)	REG
KI, RO, MU	2, 3	(2, 1, 5)	5
KI, RO, BU	3, 4	(3, 1, 4)	4
RO, MU, BU	3, 5	(3, 2, 3)	3
KI, MU, BU	3	(3, 5)	5

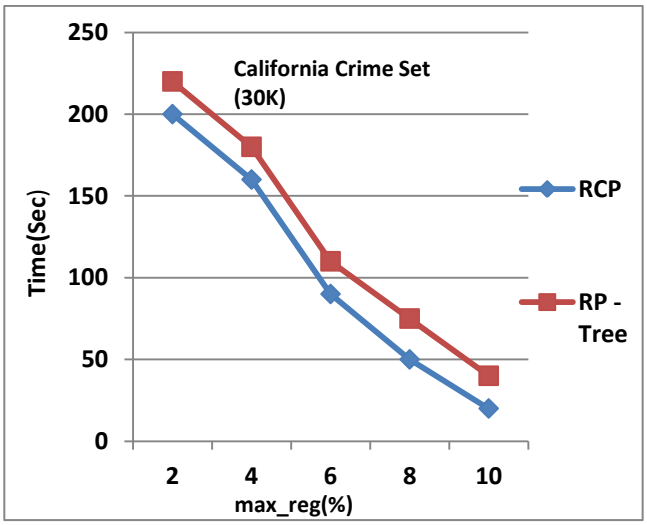
3. EXPERIMENT RESULTS

Our experimentation results are performed over crime dataset i.e., California crime dataset which is available as open source. Our algorithm RCP compared with the results of RP-tree which shows our algorithm is more efficient and fast in finding the regular crime patterns. All experiments are done in java on windows XP containing 2.7 GH with 2GB of main memory.

The execution time over California crime dataset on 50K records can be seen in Figure 1 and 30K records in Figure 2. The two figures show our algorithm is more efficient than the existing RP-tree.



[Figure 1]



[Figure 2]

4. CONCLUSION

In this paper we considered the regular patterns from crime database using vertical data format. Vertical data format approach achieves better performance it requiring only one database scan. We use this method on crime database on maximum regularity threshold.

5. REFERENCES

[1] S.K. Tanbeer, C.F. Ahmed, B.S. Jeong. "Mining regular patterns in data streams." In: DASFAA. Volume 5981 of LNCS. Springer 2010, pp. 399-413.

[2] S. Sivaranjani, S. Sivakumari, A Novel Approach for serial crime detection with the consideration of class imbalance problem, Indjst, volume 8, issue 34, December 2015.

[3] M.G. Elfeky, W.G. Aref, A.K. Elmagarmid "Periodicity detection in time series databases", IEEE Transactions on Knowledge and Data Engineering 17(7), pp. 875-887, 2005.

[4] G. Lee, W. Yang, J-M Lee. "A Parallel algorithm for mining partial periodic patterns." Information Society 176, pp. 2006, pp.3591-3609

[5] R. Agarwal, and R. Srikanth, "Fast algorithms for mining

association rules in Large databases", In Proc. 1994 Int. Conf. Very Large Databases VLDBA'94, Santiago, Chile, Sept. 1994, pp. 487- 499.

[6] B. Ozden, S. Ramaswamy, A. Silberschatz. "Cyclic Association Rules." In.: 14 th International conference on Data Engineering, 1998, pp. 412-421

[7] J. Han, J. Pie, Y. Yin "Mining Frequent Patterns without candidate generation", In Proc. ACM SIGMOD international Conference on management of Data, 2000, pp.1-12.

[8] Han J, Pie J, Yin Y "Mining Frequent Patterns without candidate generation", In Proc. ACM SIGMOD international Conference on management of Data, 2000, pp. 1-12.