SWPC 2017 Submission

Title	New Rare Pattern Mining Algorithm for Streaming Data
Author Names	Mayuresh Gharpure, Makarand Muley
WWIDs	11528336, 11528335

Summary

Rare pattern mining refers to hunting for unusual patterns from a data set. Most data mining systems are targeted to finding frequent patterns to understand the trends. However, in some domains we need to find out the rare patterns to identify an anomaly in the system. This abstract describes a new algorithm for finding rare patterns in a data stream. The algorithm described here is an improvement over the previously known best algorithm in data stream rare pattern mining. The algorithm reduces the memory ,and time required to find a rare pattern compared to current state of the art algorithm

Opportunity/Problem

With the advent of Internet of Things, the use of different sensors has increased significantly across various domains like smart homes, industrial automation and healthcare. These sensors generate a continuous stream of data which is generally sent to the cloud for analysis done periodically. Howver, in some scenarios, having intelligence at the edge can help in real time discovery of unusual patterns. E.g in an industrial plant, if many sensors operating at different stages of assembly line, could detect an anamolus pattern in the sensed data, it could potentially save significant money or even life

This solution is applicable to all such scenarios where a stream of data is available and there's a need for edge analytics either due to either performance constrains or privacy constraints associated with sending data to outside cloud

Solution

The previously known best algorithm on which we have proposed an improvement used following data structures and concepts

1) SRP Tree (Streming Rare Pattern Tree)

This is a data structure similar to the FP Tree used in frequent pattern mining algorithms, with following differences:

a) FP Tree generation needs two-passes over the data, and only items which are frequent are inserted in the tree. SRP Tree needs a single pass and makes a tree with all the elements in transaction

b) In FP-Tree, only frequent items are added in the tree and in decreasing order of frequency, where as in SRP Tree, the all items in a transaction are added, and in the canonical order

c) A connection table is maintained, to keep track of elements co-occuring (Table 3)

2) minFreqSup and minRareSup -> minFreqSup is the threshold count for considering the pattern as frequent. A pattern is considered rare if, its support count is >=minRareSup & < minFreqSup

In the example shown in Table 1, if we assume minFreqSup = 4 and minRareSup = 2, we see that items a,d,e are frequent and c,b are rare.

Using the connection table now we find the items co-occuring with rare items

Items co-occuring with b,c->d,e

So, now mine the tree using FP Growth algorithm for b,c,d,e

Abbreviations: RF – Reordered by Frequency, RC – Reordered by canonical order, FI – Frequent Items

TID	Items	R.F	R.C	F.I
1	d,e	e,d	d,e	e,d
2	a,d,e	e,d,a	a,d,e	e,d,a
3	d,b	d,b	b,d	d
4	a,e	a,a	a,e	E,a
5	a,e,c	e,a,c	a,c,e	E,a
6	e,a,d	e,d,a	a,d,e	E,d,a
7	d	d	d	d
8	e,b	e,b	b,e	e
9	e,a,d	e,d,a	a,d,e	e,d,a
10	b,c	b,c	b,c	-

Table 1

Item	Count
a	5
b	3
c	2
d	6
e	7

Table 2

Item	Items Co-Occur
a	{((e:5),(c:1),(d:3)}
b	{(d:1),(e:1),(c:1)}
c	{((e:1)}
d	{(e:4)}

Table 3

Our Improvement:

Our approach is to alter the process a bit as follows:

1) Insert every item in the tree in canonical order as in SRP-Tree
2) When a request for mining comes, re-order the branches of tree in descending order of the frequency of the items in branch
3) In the re-ordered tree, skip a branch whose leaf node is 'frequent', because, the patterns forming up from the frequent leaf node will also be frequent, so no need to mine it for rare patterns. This pruning leads to performance improvement

Following figures depict the trees generated using SRP-Tree and Modified SRP Tree with our improvement

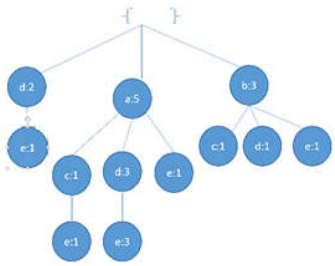

Figure 1 - SRP Tree

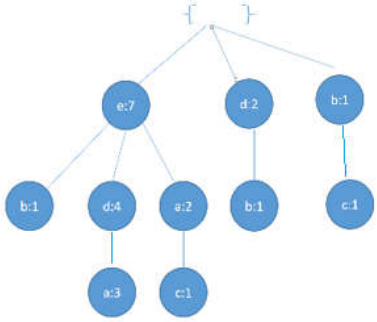

Figure 2 - Our Approach

Results:

Removed the need for connection table

Dataset	Transactions	minRareSup	minFreqSup	Window Size	Time (SRP)	Time (Our Approach)
T10I4D100K	100000	0.01%	0.05%	25000	12s	5.61s
Mushroom	8124	1%	5%	2000	1131s	179.4s

Key Takeaway

The new approach gives better performance than current state of the art algorithms and also eliminates the need for a connection table data structure.

Self-Assessment

Criteria	Supporting Details
Scope	Emerging Programming Models
Innovation	Improvement to the current state of the art algorithm for rare pattern mining
Result	Improved performance with less memory requirement
Interest	Wide scope of application from IoTG to TMG, wherever sensors generate stream of data

References

Huang D., Koh Y.S., Dobbie G. (2012) Rare Pattern Mining on Data Streams. In: Cuzzocrea A.,Dayal U. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2012. Lecture Notes in computer Science, vol 7448. Springer, Berlin, Heidelberg