Predicting Rare Events In Temporal Domains

Ricardo Vilalta and Sheng Ma
ICDM 2002

Presented by: Fahad Arshad

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

Outline

- Introduction
- Event Prediction Problem
- Searching for Eventsets
- Building a Rule-base Model for Prediction
- Experiments and Results
- Conclusion

PURDUE UNIVERSITY

Introduction

- Learning to predict rare events (*target events*) is a difficult problem
  - *Attack on a network, fraudulent transactions at bank etc*

- Prediction strategy
  1. Characterize target events by finding the types of events frequently preceding target events within fixed time window *W*
  2. Validate that these event types uniquely characterize target events and do not occur often far from target events
  3. Combine validated event types using association rule mining to build a rule-based system for prediction

PURDUE UNIVERSITY

Introduction

- Association Rule Mining
  - Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of *m* binary attributes called *items* and $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions called the *database*.
  - A *rule* is defined as an implication of the form X→Y where X,Y (*itemsets*) are subsets of *I* and X ∩ Y = *empty*
  - In the real world example, a *rule*: {milk, bread} →{butter}
  - Support(*X*) is the proportion of transactions in the dataset which contain the itemset X
  - Confidence of a rule is:

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

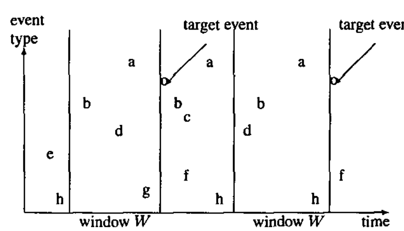
PURDUE UNIVERSITY

Event Prediction Problem

- Definition 1:
  - A sequence of events is an ordered collection of events $D = < d_1, d_2, \ldots, d_n >$, where each event d_i is a pair $d_i = (e_i, t_i)$. e_i indicates the event type and t_i its occurrence time.
- D_{target} (target events) a subset of D has size m and $m << n$
- Definition 2:
  - An eventset Z is a set of event types $\{e_i\}$. Event set Z matches the set of events in window W, if every event type $e_i \varepsilon Z$ is found in W

Event Prediction Problem

- Definition 3:
  - Eventset Z has *support* s in D if $s\%$ of all windows of size W preceding target events are matched by Z. Eventset Z is *frequent* if s is above a minimum user defined threshold
- Definition 4:
  - Eventset Z has *confidence* c in D if $c\%$ of all time windows of size W matched by Z precede a target event. Eventset Z is *accurate* if c is above a minimum user defined threshold



Searching for Eventsets

- Frequent Eventsets
  - Maintain in memory events within a sliding window of size W
  - With each new target event, all event types within W are stored in a new transaction
  - A-priori algorithm is used to find all eventsets above a minimum user defined threshold.
  - Order of events and inter-arrival times between events within each W are not relevant.

Algorithm 1: Finding Frequent Eventsets
Input: event sequence D, window size W, minimum support $s\%$, target-event type e^*
Output: frequent eventsets $\mathcal{F}$

FREQUENTEVENTSETS(D,W,s,e^*)
$B = \emptyset$; $T = \emptyset$
foreach event $d_i = (e_i, t_i) \in D$
  currentTime $= t_i$
  foreach event $d_j = (e_j, t_j) \in T$
    if (currentTime $- t_j) > W$
      Remove d_j from T
  end
  if d_i is a target event (i.e., $e_i = e^*$)
    $B = B \cup \{e_j \mid (e_j, \cdot) \in T\}$
  $T = T \cup d_i$
end
Use A-priori over B to find all frequent
eventsets with minimum support $s\%$.
Let $\mathcal{F}$ be the set of all frequent eventsets.
return $\mathcal{F}$

Searching for Eventsets

- Accurate Eventsets
  - Find the number of times that each of the eventsets occur outside W preceding target events
  - Compute the confidence of each frequent eventset and eliminate those below the threshold
  - B' a database of all eventsets not preceding target events is found.
  - If x_1 and x_2 are the number of transaction in B and B' respectively
  - confidence(Z, B, B') = $x_1/(x_1+x_2)$

Algorithm 2: Finding Confident Eventsets
Input: event sequence D, minimum confidence $c\%$, time intervals I, frequent eventsets $\mathcal{F}$, database eventsets B
Output: confident eventsets $\mathcal{F}'$

CONFIDENTEVENTSETS(D,c,I,$\mathcal{F}$,B)
$T = \emptyset$; [a,b] = next interval from I
foreach $d_i = (e_i, t_i) \in D$
  if $t_i \in [a, b]$
    $T = T \cup d_i$
  if $t_i > b$
    $B' = B' \cup \{e_j \mid (e_j, \cdot) \in T\}$
    $T = \emptyset$; $[a, b]$ = next interval from I
    Add d_i to D
end
$\mathcal{F}' = \emptyset$
foreach eventset Z in $\mathcal{F}$
  if confidence(Z,B,B')> c AND
  $P(Z|B) > P(Z|B')$
    $\mathcal{F}' = \mathcal{F}' \cup Z$
end
return $\mathcal{F}'$

Searching for Eventsets

- Validation step for Accurate Eventset:
  - Let P(Z|B) be the probability of Z occurring within database B and P(Z|B') the corresponding probability within B'
  - Event Z is validated if we reject the null hypothesis H_0 with high confidence

$$H_0 : P(Z|B) \leq P(Z|B')$$

  - If number of events is large, one assume a Gaussian distribution and reject the null hypothesis in favor of the alternate hypothesis H_1

$$H_1 : P(Z|B) > P(Z|B')$$

  - For a given confidence level α, if the difference between the two probability is significant, we reject H_0. By choosing a small α we can almost be certain about the relation of Z with target events.

Building a Rule-base Model

- Definition 5:
  - Eventset Z_i is said to be more specific than eventset Z_j, if Z_j is a subset of Z_i
    - e.g.{a, b, c} is more specific than {a, b}

- Definition 6:
  - Eventset Z_i is said to have higher rank over eventset Z_j, represented as $Z_i > Z_j$ if any of the following conditions is true

    1. The confidence of Z_i is greater than that of Z_j
    2. The confidence of Z_i equals that of Z_j, but the support of Z_i is greater than that of Z_j
    3. The confidence and support of Z_i equal that Z_j, but Z_i is more specific than Z_j

Building a Rule-base Model

- Find the most accurate and specific rule first
- Rule-based system *R* can be used for prediction by checking the occurrence of any of the events in *R* along the event sequence used for testing.
- The model predicts finding a target event within a time window of size *W* after any such eventset is detected.

Algorithm 3: Building Rule-Based Model
Input: eventsets $\mathcal{F}'$
Output: Set of rules $\mathcal{R}$

```
RULE-BASED-EVENTSETS(F')
    R = ∅
    Sort F' in decreasing order by rank
    while F' is not empty
        Let Z_i be the first eventset in F'
        if Z_j ⊂ Z_i, i ≠ j, remove Z_j from F'
        Make a new rule r : Z_i → targetevent
        R = R ∪ r
        Remove Z_i from F'
    end
    return R
```


Experiments and Results

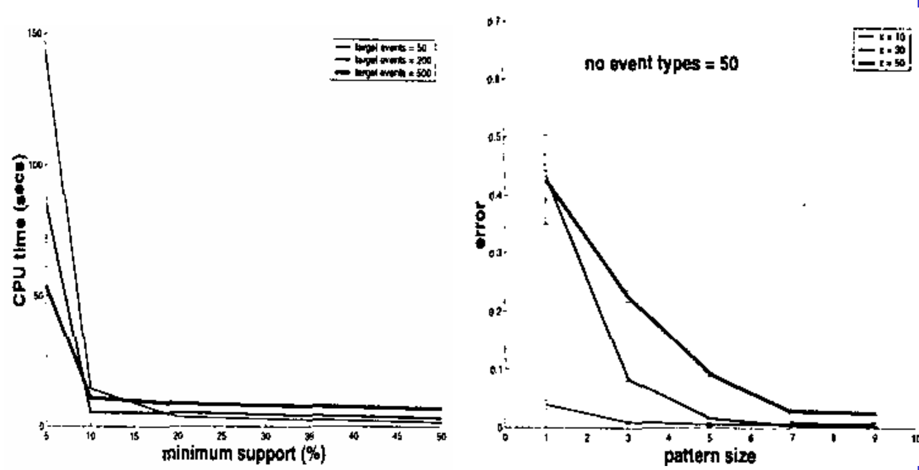
- Artificial Data
  - Data generator outputs uniformly distributed sequence of events over a fixed time interval
  - Time-interval= 1 week, number of event types = 50, W = 5 min, target events = 50, s = 0.1, significance level for hypothesis testing α = 0.01.
  - For each event sequence first 50% is for training and the rest for testing. Results are averaged over 30 runs.

Experiments and Results

- Non-overlapping windows of size W that do not intersect the set of time windows preceding target events are negative examples.
- All time windows preceding target events are considered positive examples
- *Error* is the fraction of examples incorrectly classified by the rule –based system.

PURDUE UNIVERSITY

Experiments and Results

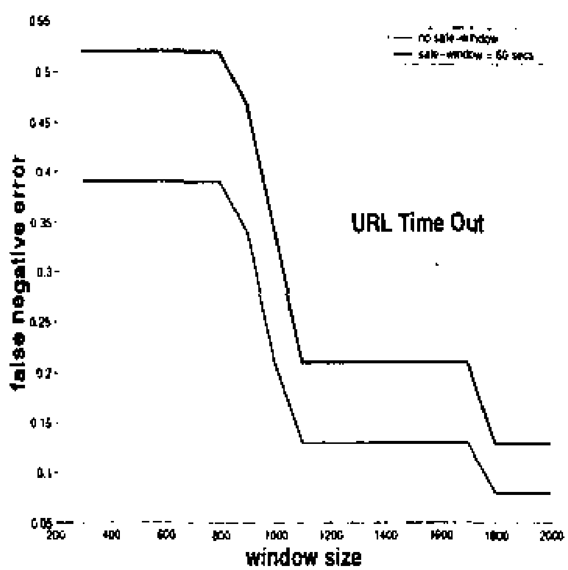


PURDUE UNIVERSITY

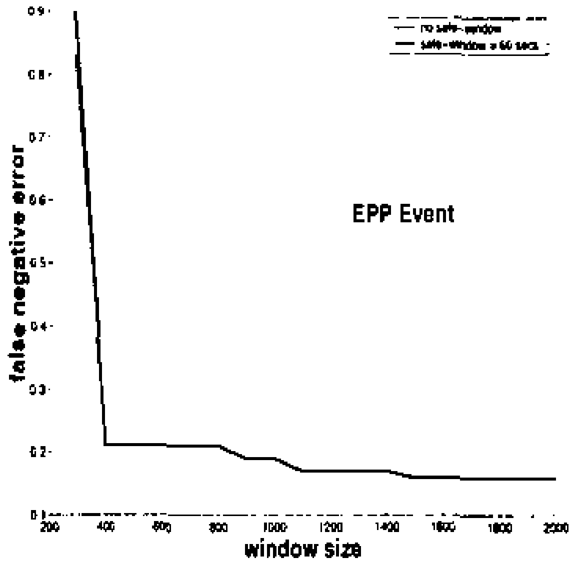
Experiments and Results

- Real Production Data:
  - Data collected over 1 month on a network with 750 hosts
  - 26000 events, 165 different types of events
- Two types of target events considered
  - EPP event indicates that end-to-end response time for a host is above a critical threshold
  - URL timeout indicates that a website is inaccessible.
- Event is characterized by time, event type, and host
  - Merge type and host to identify the nature of the event

Experiments and Results



Experiments and Results



PURDUE UNIVERSITY

Conclusions

- An approach to detecting patterns in events sequences before a target event.
- Size of the time window preceding target events is crucial to this approach
- Two different combination of event-types and host of interest show how the false negative rate decreases significantly with increase in window size *W*
- The approach is contingent on sets of events frequently occurring before target events.

PURDUE UNIVERSITY