Classification Rule Mining for a Stream of Perennial Objects

Zaigham Faraz Siddiqui and Myra Spliopoulou

Otto-von-Guericke Universitaet

RuleML 2011, Barcelona, Spain.

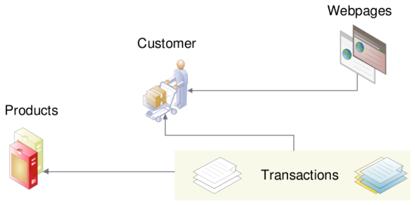
Overview

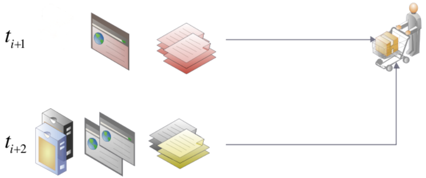
Conventional Stream Mining

- Objects come from single stream
- Objects arrive as an ordered sequence
    - $x_1, x_2, \ldots, x_i, \ldots$
- Objects are independent of each other
    - Unique identifiers
- Objects are static
    - can be forgotten when they grow old

Stream of Perennial Objects

- What are perennial objects?
  - are from a relational stream
    - are linked to other objects
  - are dynamic
    - can change their definition
    - can change their class label
  - cannot be forgotten



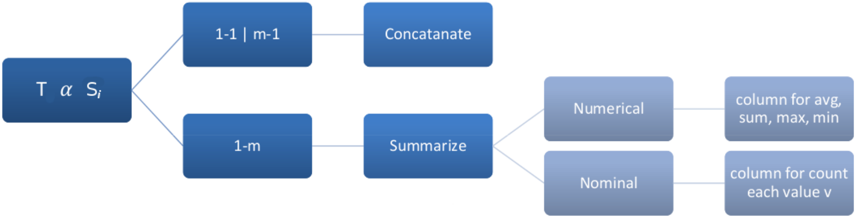


Mining Perennial Objects

- Determine the perennial stream T
  - Customer, Accounts, etc.
- A perennial object may be linked more than one object in the neighbouring streams that provides an extensional definition for a perennial object.
- Traditional relational algorithms are limited to static mining and are usually very expensive.
- The two approaches that work over streams SRPT & TrIP use aggregated or summarised information to build their model.

Incremental Propositionalisation

- Memory Management
  - Windows for fast/ephemeral streams.
  - Caches for slower/perennial streams.



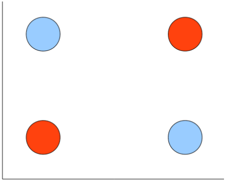
CID	Age	Gender
1	50	M
2	24	F

PID	CID	Color	Price
1	1	R	500
2	1	B	40
3	2	G	5
4	2	G	10

CID	Age	Gender	Count Color			Summ Price		
			R	B	G	Min	Max	Avg
1	50	M	1	1	0	40	500	270
2	24	F	0	0	2	5	10	7.5

Motivation

- Aggregates make very simplistic assumptions
  - Every attributes is independent
  - Numerical attributes are normally distributed



- Propositionalisation generates too many aggregates
  - Information content in aggregates is low
  - Results in very deep decision trees

Basics

- Rules are are learned on fast streams
- Labels are propagated from the perennial stream
- Rules are stored in a concept lattice $\mathcal{L}$
- Each rule $\mathcal{I}$ has:
  - the form $\mathcal{I} : X \wedge Y \rightarrow [p_1, \ldots, p_l]$
    e.g.,$A_{red} \wedge B_{big} \rightarrow$[$10^{+}$,$90^{-}$]
  - a creation timepoint
  - a lower bound on the missed examples

Rule Discoverey: CRMPES Algorithm

1 *increment* $\mathcal{L}$ for new tuples in S_j

2 *grow* $\mathcal{L}$ once all tuples have arrived

Grow Lattice

- $\mathcal{L}$ is grown pro-actively
- a new rule $\mathcal{I}'$ is added to $\mathcal{L}$, if
  - its parents have min support
  - its parents are not locked
- such rule is marked as *tentative*

3 *decrement* $\mathcal{L}$ for old tuples in S_j

4 *shrink* $\mathcal{L}$

Shrink Lattice

- Remove *redundant* rules
- *Lock* rules that offers no improvement
- Lock rules that do not meet the min support
  - delete its children
  - re-introduce redundant rules

Using rules to generate new features

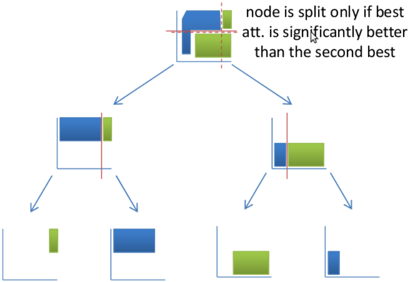
- Rules are first ranked for uninterestingness

$$d(\mathcal{I}) = support(\mathcal{I}) \times e(\mathcal{I})$$

- Top ranked rule from list $\mathcal{R}$ is selected and added to the set $\mathcal{F}$
- $\mathcal{R}$ is traversed and the rules with min intersection with $\mathcal{F}$ are incorporated into $\mathcal{F}$
- Termination condition: either $\mathcal{F}$ or rule list exhausts
- Antecedent of each rule in $\mathcal{F}$ is converted into an attribute

Decision tree Induction

- Initialize root node R
- for $i = 1 \rightarrow END$
  - $\mathcal{L} \leftarrow \text{CRMPES}(\mathcal{L}, \mathcal{X}_i)$
  - $\mathcal{F} \leftarrow \text{FGen}(\mathcal{L}, \mathcal{F})$
  - $\mathcal{W}_i \leftarrow \text{IncProp}(\mathcal{X}_i)$
  - $\zeta \leftarrow \text{AdaptDecisionTree}(\zeta, \mathcal{W}_i, \mathcal{F})$
- The node splits are decided using Hoeffding bound



Evaluation Setting

Dataset

- Sythetic dataset: simulates users' buying behaviour for different items.
- Financial dataset: accounts information for a period of years

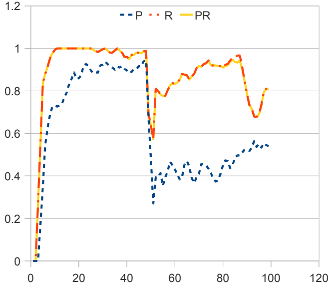
Strategies

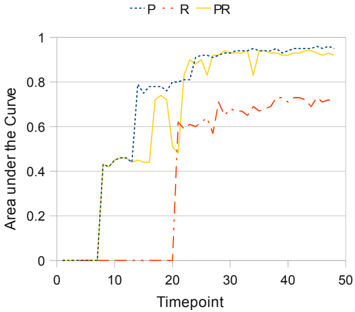
- P: uses simple aggregates only.
- R: uses rule-based attribute.
- PR: uses simple aggregates and rule-based attributes.

Objectives

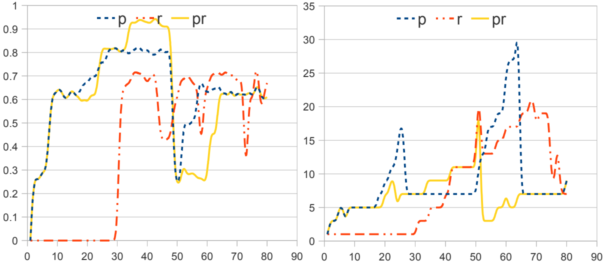
- The performance of strategies P, R and PR with respect to the information content they hold.

Learning User Profiles



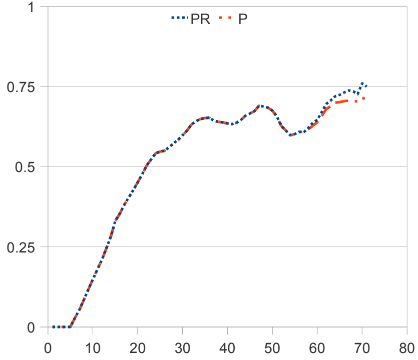


Learning User Profiles



Learning Bank Accounts

- Labels become applicable once the accounts have matured
- Rules with predictive power were discovered quite later



Summary and Future Work

Summary

- Algorithms for perennial objects are scarce
- Uses classification rules to enhance the tree induction
- Results in shorter trees with better performance

Future Work

- More experiments on real dataset required
- Classification rules can help reduce the number of generated features

Questions?