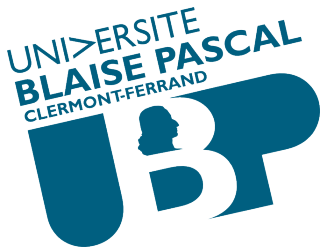
A tool for classification of sequential data

Giacomo Kahn, Yannick Loiseau, Olivier Raynaud

LIMOS

August 30, 2016



AUVERGNE – RhôneAlpes

Classifier

Tool

Dataset

Experimental results

Problem

C_M dataset :	C_S dataset :
C_M ; event 1	C_S ; event 1
C_M ; event 2	C_S ; event 2
...	...
C_M ; event n	C_S ; event m

Given a subset of events from a class, we want to classify this subset.

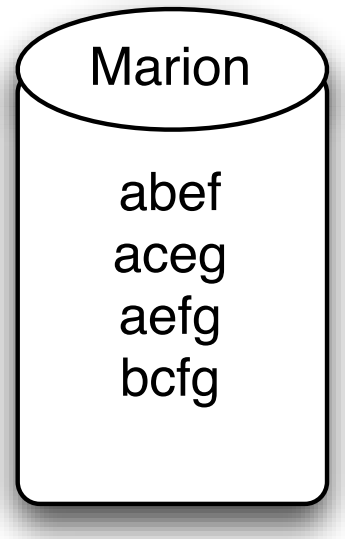
Existing solutions

- Bayesian procedures (statistic procedures)
- Support Vector Machines (linear classifier)
- FCA : based on negative and positive examples
- FCA : based on pattern structures
- . . .

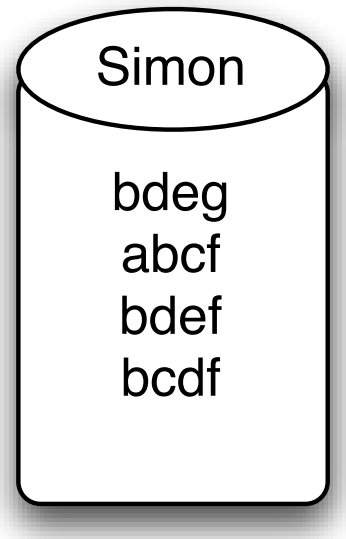
Context of the study

Context

The context of our study is implicit authentification from web-navigation related events. This forces us to group events into sessions.

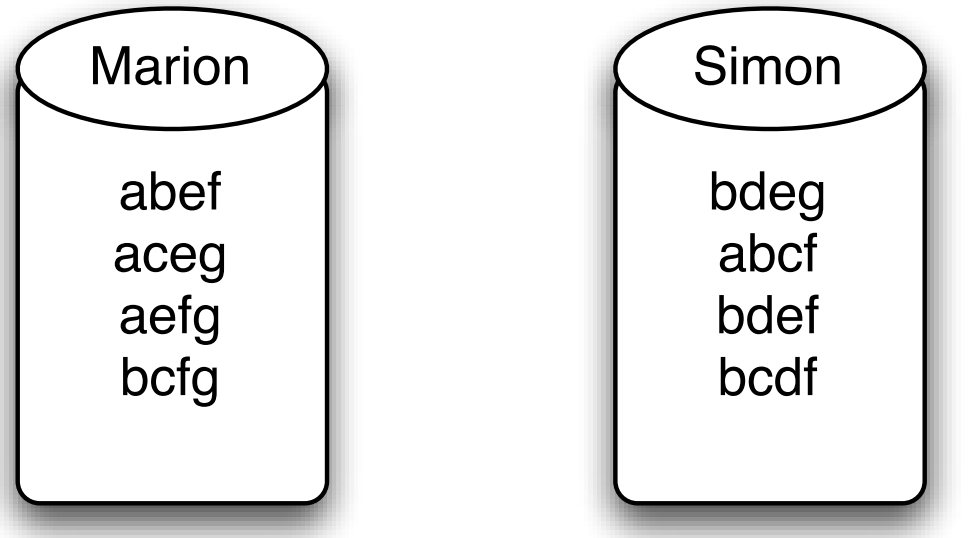


Marion's sessions



Simon's sessions

Implemented method



Marion's sessions

Simon's sessions

Patterns

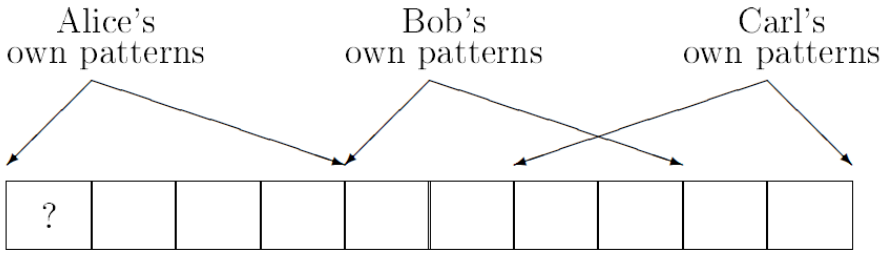
Marion's *own patterns* $= \{ae, ag, fg, aeg, aef\}$

Simon's *own patterns* $= \{bdf, bde, abcf, bcdf\}$

Implemented method

Common vector

We compute a common vector from all *own patterns*.



V_{Marion}

3/4	1/2	1/2	1/2	0	0	1/2	0	0	0
ae	ag	fg	aeg	bdf	bde	aef	$abcf$	$bcdf$	$abdf$

V_{Simon}

0	0	0	0	1/2	1/2	0	1/4	1/4	0
ae	ag	fg	aeg	bdf	bde	aef	$abcf$	$bcdf$	$abdf$

Example

Anonymous dataset *A*

A own patterns $= \{ae, ag, fg, aeg, aef, bcdf\}$

$V_{Anonymous}$

1/2	3/4	3/4	1/2	0	0	1/2	0	1/2	0
ae	ag	fg	aeg	bdf	bde	aef	$abcf$	$bcdf$	$abdf$

Similarity measure

$V_{Anonymous}$ is more similar to V_{Marion} than to V_{Simon}. We conclude that the anonymous dataset *A* is from the class *Marion*.

Classifier

Tool

Dataset

Experimental results

Parameters

Wide range of parameters for the classifier

- Constitution of sessions (timestamps, fixed-size, time-delay...)
- Metrics used (support, lift, $tf \times idf$)
- Fuzzy inclusion for anonymous patterns
- Various similarity measures (cosine, Kulczynski)

For the experiments

- Number of runs
- Number of anonymous sessions received
- Classification mode (binary or non-binary)

Classifier

Tool

Dataset

Experimental results

cez13 dataset

Our case study of implicit authentification is supported by a dataset of connection logs from Blaise Pascal university servers.

Data

- Raw data : 17×10^6 lines in log format
- Filtered data : 4×10^6 lines
- 3000 classes (we used the 150 more frequent for our study)

Duration

The data has been collected over a 6-month period during the school year 2012-2013.

Filters

Black list filters

- `http://winhelp2002.mvps.org/hosts.htm`
- `https://pgl.yoyo.org/as`

HTTP requests filters

We used the status code from simple HTTP request to keep the websites that are more relevant to the study.

	#Users	*#Sites*	Avg#lines/user
Raw Data	3388	96184	5082
Filtered data	3370	57654	1145

Classifier

Tool

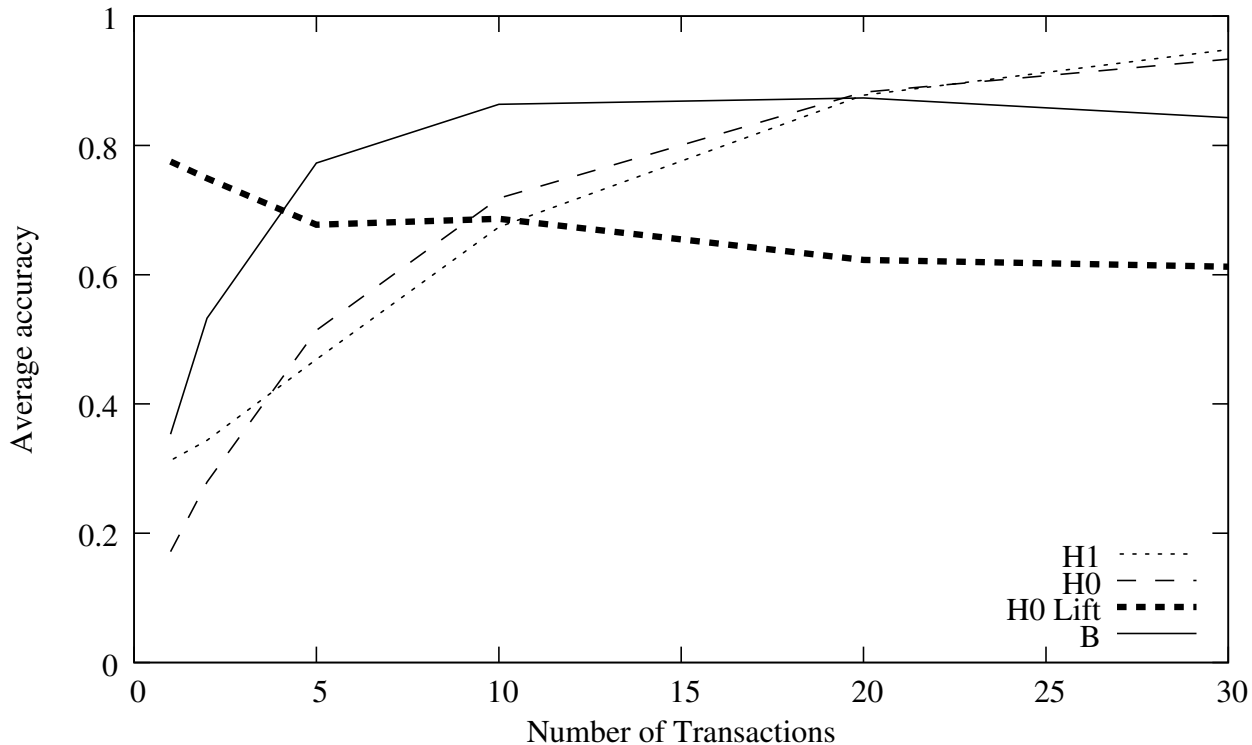
Dataset

Experimental results

Sequential data

Cez13 (150 classes)

Method	Accuracy	Ratio classified/not classified
Bayes	0.35 - 0.84	1
H1	0.31 - 0.95	0.75 - 1



Adaptation for a traditional classification task

p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	n	g
e	b	s	w	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	n	m
e	b	y	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	s	m
p	x	y	w	t	p	f	c	n	p	e	e	s	s	w	w	p	w	o	p	k	v	g
e	b	s	y	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	s	m
e	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n	n	g
e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	s	m
e	b	s	y	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n	s	g
p	x	y	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	v	u
e	x	f	n	f	n	f	w	b	n	t	e	s	f	w	w	p	w	o	e	k	a	g
e	s	f	g	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	y	u

Data formating

Change in the data

- 1 tuple = 1 session
- session size = number of properties of the dataset

Pattern considered

In this case, we only consider fixed-size sessions and closed-itemsets patterns.

Data formating

```
p; 1000; cap-shape=p
p; 1000; cap-surface=x
p; 1000; cap-color=s
p; 1000; bruise=n
p; 1000; odor=t
p; 1000; gill-attach=p
p; 1000; gill-spacing=f
p; 1000; gill-size=c
p; 1000; gill-color=n
p; 1000; stalk-shape=k
p; 1000; stalk-root=e
p; 1000; stalk-surface-above=e
p; 1000; stalk-surface-below=s
p; 1000; stalk-color-above=s
p; 1000; stalk-color-below=w
p; 1000; veil-type=w
p; 1000; veil-color=p
p; 1000; ring-number=w
p; 1000; ring-type=o
p; 1000; spore-print-color=p
p; 1000; population=k
p; 1000; habitat=s
p; 1003; cap-shape=p
p; 1003; cap-surface=x
p; 1003; cap-color=y
p; 1003; bruise=w
p; 1003; odor=t
p; 1003; gill-attach=p
```

Traditional classification task (work in progress)

Breast-wisconsin (binary)

Method	Accuracy
Bayes	0.97
LLO	0.94
H1	**0.98**

Mushroom (binary)

Method	Accuracy
Bayes	0.99
SVM	0.91
H1	**1**

Bayes : smoothed Bayes classifier, LLO : Lazy Lattice-based Optimization, SVM : Support Vector Machine

Conclusion

- Tool for sequential data
- A new dataset available at `http://fc.isima.fr/~kahngi/cez13.zip`
- Works fine on traditional classification tasks

Perspectives

- Improve the sequence mining
- Other types of patterns (pattern structures, association rules)
- Aggregations operators for fuzzy inclusion
- Include some clustering from the profiles vectors