Sequence Classification Using Both Positive & Negative Patterns and Its Application for Debt Detection

Yanchang Zhao[1], Huaifeng Zhang[2], Shanshan Wu[1], Jian Pei[3], Longbing Cao[1], Chengqi Zhang[1], and Hans Bohlscheid[2]

[1] University of Technology, Sydney, Australia
[2] Centrelink, Australia
[3] Simon Fraser University, Canada

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

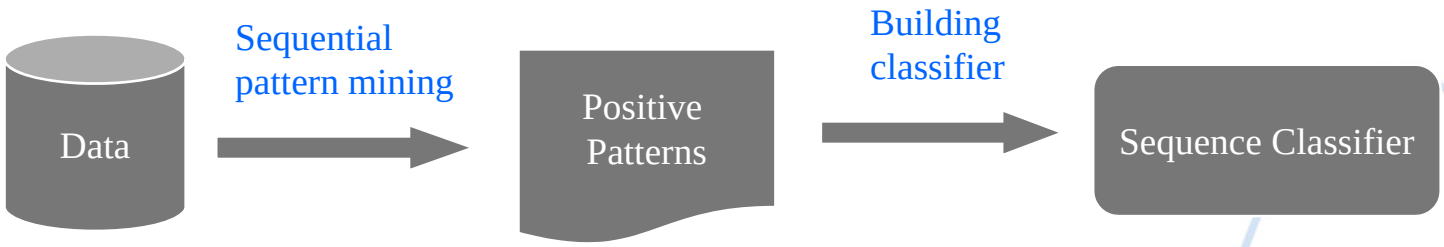
17 December 2009

Contents

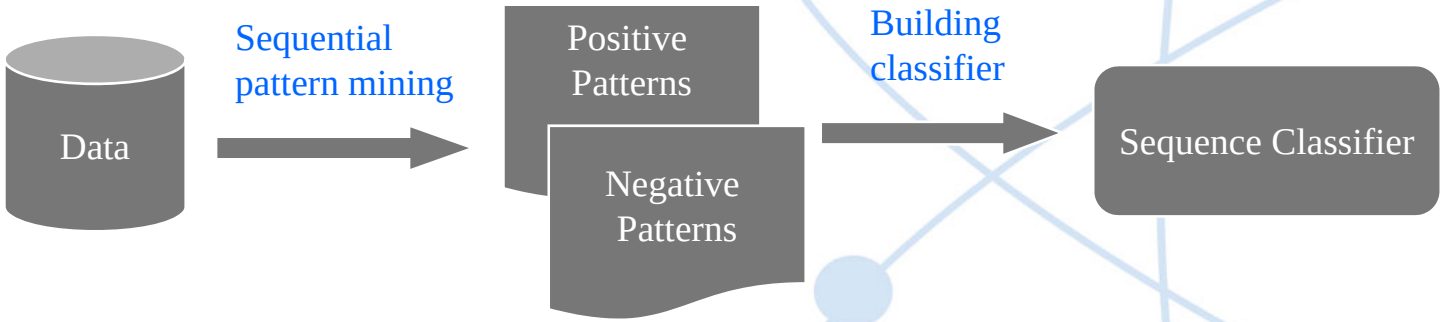
- Introduction
- Related Work
- Sequence Classification Using Both Positive and Negative Patterns
- Experimental Evaluation
- Conclusions

Sequence Classification

Traditional way: using positive patterns only



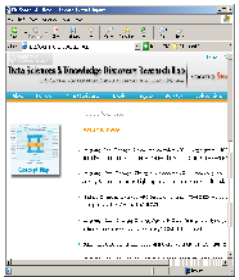
Proposed technique: using both positive and negative patterns



Negative Sequential Patterns

- Positive sequential patterns
  - ABC
- Negative sequential patterns: sequential patterns with the non-occurrence of some items
  - AB(¬D)
- Negative sequential rules
  - AB → ¬D
  - ¬(AB) → D
  - ¬(AB) → ¬D

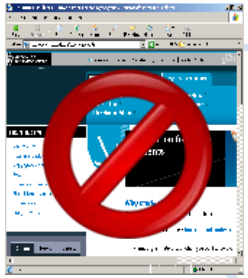
An Example



Booking

Joining

Buying book A



No booking

No joining

Buying book B

Related Work

- Negative Sequential Patterns
- Sequence Classification
- Fraud/Intrusion Detection

Positive Sequential Pattern Mining

- GSP (Generalized Sequential Patterns), Srikant & Agrawal, EDBT'96
- FreeSpan, Han et al., KDD'00
- SPADE, Zaki, Machine Learning 2001
- PrefixSpan, Pei et al., ICDE'01
- SPAM, Ayres et al., KDD'03

Only positive patterns are considered.

Negative Sequential Patterns

- Sun et al., PAKDD' 04: $\neg P \xrightarrow{T} e$
- Bannai et al. WABI'04: $p' \wedge q'$ and $p' \vee q'$ where p' is either p or $\neg p$.
- Ouyang and Huang, ICMLC' 07: $(A, \neg B)$, $(\neg A, B)$ and $(\neg A, \neg B)$
- Lin et al. ICACS'07: only last item can be negative
- Zhao et al., WI'08, PAKDD'09: Impact-oriented negative sequential rules

Sequence Classification

- Lesh et al., KDD'99: using sequential patterns as features to build classifiers with stand classification algorithms, such as Naïve Bayes.
- Tseng and Lee, SDM'05: Algorithm CBS (Classify-By-Sequence). Sequential pattern mining and probabilistic induction are integrated for efficient extraction of sequential patterns and accurate classification.
- Li and Sleep, ICTAI'05: using n-grams and Support Vector Machine (SVM) to build classifier.
- Yakhnenko et al., ICDM'05: A discriminatively trained Markov Model (MM(k-1)) for sequence classification
- Xing et al., SDM'08: early prediction using sequence classifiers.

Negative sequential patterns are NOT involved.

Fraud/Intrusion Detection

- Bonchi et al., KDD'99: using decision tree (C5.0) for planning audit strategies in fraud detection
- Rosset et al., KDD'99: fraud detection in telecommunication, base on C4.5
- Julisch & Dacier, KDD'02: using episode rules and conceptual classification for network intrusion detection
- …

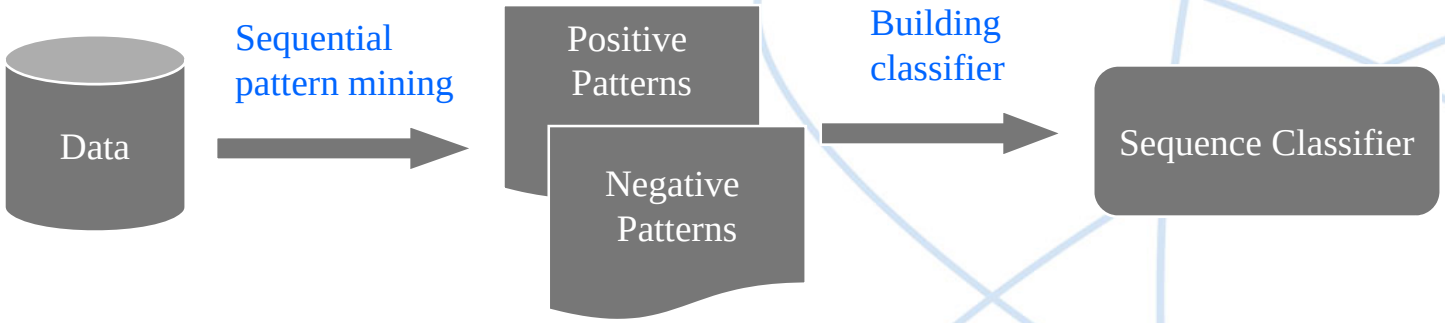
Negative sequential patterns are NOT involved.

Contents

- Introduction
- Related Work
- Sequence Classification Using Both Positive and Negative Patterns
- Experimental Evaluation
- Conclusions

Problem Statement

- Given a database of sequences, find all both positive and negative discriminative sequential rules and use them to build classifiers



Negative Sequential Rules

- Type I : $A \rightarrow B$, which means that if A appears, then B will occur;
- Type II: $A \rightarrow \neg B$, which means that pattern A is not followed by B;
- Type III: $\neg A \rightarrow B$, which means that if A doesnot appear, then B will occur; and
- Type IV: $\neg A \rightarrow \neg B$, which means that if A doesnot appear, then B will not occur.

Supports, Confidences and Lifts

Table 1. Supports, Confidences and Lifts of Four Types of Sequential Rules

Type	Rules	Support	Confidence	Lift
I	$A \rightarrow B$	$P(AB)$	$\frac{P(AB)}{P(A)}$	$\frac{P(AB)}{P(A)P(B)}$
II	$A \rightarrow \neg B$	$P(A)-P(AB)$	$\frac{P(A)-P(AB)}{P(A)}$	$\frac{P(A)-P(AB)}{P(A)(1-P(B))}$
III	$\neg A \rightarrow B$	$P(B)-P(A\&B)$	$\frac{P(B)-P(A\&B)}{1-P(A)}$	$\frac{P(B)-P(A\&B)}{P(B)(1-P(A))}$
IV	$\neg A \rightarrow \neg B$	$1-P(A)-P(B)+P(A\&B)$	$\frac{1-P(A)-P(B)+P(A\&B)}{1-P(A)}$	$\frac{1-P(A)-P(B)+P(A\&B)}{(1-P(A))(1-P(B))}$

- A&B: A and B appears in a sequence
- AB: A followed by B in a sequence
- P(A&B) >= P(AB)

Sequence Classifier

- Sequence classifier:

$$\mathcal{F} : \mathcal{S} \xrightarrow{\mathcal{P}} \mathcal{T},$$

where S is a sequence dataset, T is the target class, and P is a set of classifiable sequential patterns (including both positive and negative ones).

Discriminative Sequential Patterns

- CCR (Class Correlation Ratio), Verhein & Chawla, ICDM'07:

$$CCR(p_a \rightarrow \tau) = \frac{\hat{corr}(p_a \rightarrow \tau)}{\hat{corr}(p_a \rightarrow \neg\tau)} = \frac{a \cdot (c+d)}{c \cdot (a+b)},$$

Table 2. Feature-Class Contingency Table

	p_a	$\neg p_a$	$\sum$
τ	a	b	$a+b$
$\neg\tau$	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$n = a+b+c+d$

Discriminative Sequential Patterns

- The patterns are ranked and selected according to their capability to make correct classification.

$$W_s = \begin{cases} CCR, & \text{if } CCR \geq 1 \\ \frac{1}{CCR}, & \text{if } 0 < CCR < 1 \\ M, & \text{if } CCR = 0 \end{cases},$$

where M is the maximum W_s of all rules where $CCR \neq 0$.

Building Sequence Classifier

1) Finding negative and positive sequential patterns (Zhao et al., PAKDD'09);
2) Calculating the chi-square and CCR of every classifiable sequential pattern, and only those patterns meeting *support*, *significance* (measured by chi-square) and *CCR* criteria are kept;
3) Pruning patterns according to their *CCR*s (Li et al., ICDM'01);
4) Conducting serial coverage test. The patterns which can correctly cover one or more training samples in the test are kept for building a sequence classifier;
5) Ranking selected patterns with *Ws* and building the classifier. Given a sequence instance *s*, all the classifiable sequential patterns covering *s* are extracted. The sum of the weighted score corresponding to each target class is computed and then *s* is assigned with the class label corresponding to the largest sum.

Contents

- Introduction
- Related Work
- Sequence Classification Using Both Positive and Negative Patterns
- Experimental Evaluation
- Conclusions

Data

- The debt and activity transactions of 10,069 Centrelink customers from July 2007 to February 2008.
- There are 155 different activity codes in the sequences.
- After data cleaning and preprocessing, there are 15,931 sequences constructed with 849,831 activities.

Examples of Activity Transaction Data

Person_ID	Activity_Code	Activity_Date	Activity_Time
******002	DOC	20/08/2007	14:24:13
******002	RPT	20/08/2007	14:33:55
******002	DOC	05/09/2007	10:13:47
******002	ADD	06/09/2007	13:57:44
******002	RPR	12/09/2007	13:08:27
******002	ADV	17/09/2007	10:10:28
******002	REA	09/10/2007	07:38:48
******002	DOC	11/10/2007	08:34:36
******002	RCV	11/10/2007	09:44:39
******002	FRV	11/10/2007	10:18:46
******002	AAI	07/02/2008	15:11:54

Sequential Pattern Mining

- Minimum support = 0.05
- 2,173,691 patterns generated
- The longest patterns: 16 activities
- 3,233,871 sequential rules, including both positive and negative ones

Selected Positive and Negative Sequential Rules

Type	Rule	Support	Confidence	Lift
I	REA ADV ADV→DEB	0.103	0.53	2.02
	DOC DOC REA REA ANO→DEB	0.101	0.33	1.28
	RPR ANO→DEB	0.111	0.33	1.25
	RPR STM STM RPR→DEB	0.137	0.32	1.22
	MCV→DEB	0.104	0.31	1.19
	ANO→DEB	0.139	0.31	1.19
	STM PYI→DEB	0.106	0.30	1.16
II	STM PYR RPR REA RPT→ ¬DEB	0.166	0.86	1.16
	MND→ ¬DEB	0.116	0.85	1.15
	STM PYR RPR DOC RPT→ ¬DEB	0.120	0.84	1.14
	STM PYR RPR REA PLN→ ¬DEB	0.132	0.84	1.14
	REA PYR RPR RPT→ ¬DEB	0.176	0.84	1.14
	REA DOC REA CPI→ ¬DEB	0.083	0.83	1.12
	REA CRT DLY→ ¬DEB	0.091	0.83	1.12
	REA CPI→ ¬DEB	0.109	0.83	1.12
III	¬{PYR RPR REA STM}→DEB	0.169	0.33	1.26
	¬{PYR CCO}→DEB	0.165	0.32	1.24
	¬{STM RPR REA RPT}→DEB	0.184	0.29	1.13
	¬{RPT RPR REA RPT}→DEB	0.213	0.29	1.12
	¬{CCO RPT}→DEB	0.171	0.29	1.11
	¬{CCO PLN}→DEB	0.187	0.28	1.09
	¬{PLN RPT}→DEB	0.212	0.28	1.08
IV	¬{ADV REA ADV}→ ¬DEB	0.648	0.80	1.08
	¬{STM EAN}→ ¬DEB	0.651	0.79	1.07
	¬{REA EAN}→ ¬DEB	0.650	0.79	1.07
	¬{DOC FRV}→ ¬DEB	0.677	0.78	1.06
	¬{DOC DOC STM EAN}→ ¬DEB	0.673	0.78	1.06
	¬{CCO EAN}→ ¬DEB	0.681	0.78	1.05

The Number of Patterns in PS10 and PS05

	PS10 ($min\_sup = 0.1$)		PS05 ($min\_sup = 0.05$)	
	Number	Percent(%)	Number	Percent(%)
Type I	93,382	12.05	127,174	3.93
Type II	45,821	5.91	942,498	29.14
Type III	79,481	10.25	1,317,588	40.74
Type IV	556,491	71.79	846,611	26.18
Total	775,175	100	3,233,871	100

Four Pattern Sets

	Min_supp=0.10	Min_supp=0.05
Number of patterns: 4000	PS10-4K	PS05-4K
Number of patterns: 8000	PS10-8K	PS05-8K

Classification Results with Pattern Set PS05-4K

Pattern Number		40	60	80	100	150	200	300
Neg&Pos	Recall	.438	.416	.286	.281	.422	.492	.659
	Precision	.340	.352	.505	**.520**	**.503**	.474	.433
	Accuracy	.655	.670	**.757**	**.761**	**.757**	.742	.705
	Specificity	.726	.752	.909	.916	.865	.823	.720
Positive	Recall	.130	.124	.141	.135	.151	.400	.605
	Precision	.533	.523	.546	.472	.491	.490	.483
	Accuracy	.760	.758	.749	.752	.754	.752	.745
	Specificity	.963	.963	.946	.951	.949	.865	.790

In terms of recall, our classifiers outperforms traditional classifiers with only positive rules under most conditions.

Our classifiers are superior to traditional ones with 80, 100 and 150 rules in recall, accuracy and precision.

Classification Results with Pattern Set PS05-8K

Pattern Number		40	60	80	100	150	200	300
Neg&Pos	Recall	.168	.162	.205	.162	.173	.341	.557
	Precision	.620	.652	.603	.625	.615	.568	.512
	Accuracy	.771	.774	.773	.771	.771	.775	.762
	Specificity	.967	.972	.956	.969	.965	.916	.829
Positive	Recall	.141	.103	.092	.092	.108	.130	.314
	Precision	.542	.576	.548	.548	.488	.480	.513
	Accuracy	.761	.762	.760	.760	.754	.753	.760
	Specificity	.962	.976	.976	.976	.963	.955	.904

Classification Results with Pattern Set PS10-4K

Pattern Number		40	60	80	100	150
Neg&Pos	Recall	0	.303	.465	.535	.584
	Precision	0	.514	.360	.352	.362
	Accuracy	.756	.760	.667	.646	.647
	Specificity	1	.907	.733	.682	.668
Positive	Recall	.373	.319	.254	.216	.319
	Precision	.451	.421	.435	.430	.492
	Accuracy	.736	.727	.737	.738	.753
	Specificity	.853	.858	.893	.907	.893

Our best classifier is the one with 60 rules, which is better in all the three measures than traditional classifiers.

Classification Results with Pattern Set PS10-8K

Pattern Number		40	60	80	100	150	200
Neg&Pos	Recall	0	.303	.465	.535	.584	N/A
	Precision	0	.514	.360	.352	.362	N/A
	Accuracy	.756	.760	.667	.646	.647	N/A
	Specificity	1	.907	.733	.682	.668	N/A
Positive	Recall	.459	.427	.400	.378	.281	.373
	Precision	.385	.397	.430	.438	.464	.500
	Accuracy	.688	.701	.724	.729	.745	.756
	Specificity	.762	.790	.829	.843	.895	.879

Our best classifier is the one with 60 rules, which is better in all the three measures than traditional classifiers.

The Number of Patterns in the Four Pattern Sets

Pattern Set	PS10-4K	PS10-8K	PS05-4K	PS05-8K
Type I	2,621	5,430	1,539	1,573
Type II	648	1,096	2,457	6,420
Type III	2	5	0	0
Type IV	729	1,469	4	7
Total	4,000	8,000	4,000	8,000

Conclusions

- A new technique for building sequence classifiers with both positive and negative sequential patterns.
- A case study on debt detection in the domain of social security.
- Classifiers built with both positive and negative patterns outperforms classifiers built with positive ones only.

Future Work

- To use time to measure the utility of negative patterns and build sequence classifiers for early detection;
- To build an adaptive online classifier which can adapt itself to the changes in new data and can be incrementally improved based on new labelled data (e.g., new debts).

The End

Thanks!

yczhao@it.uts.edu.au

http://www-staff.it.uts.edu.au/~yczhao/