

Personal Bankruptcy Prediction Using Sequence Mining

Tengke Xiong, Shengrui Wang, André Mayers, Ernest Monga
{tengke.xiong, shengrui.wang, andre.mayers, ernest.monga}@usherbrooke.ca

University of Sherbrooke, Canada



Outline

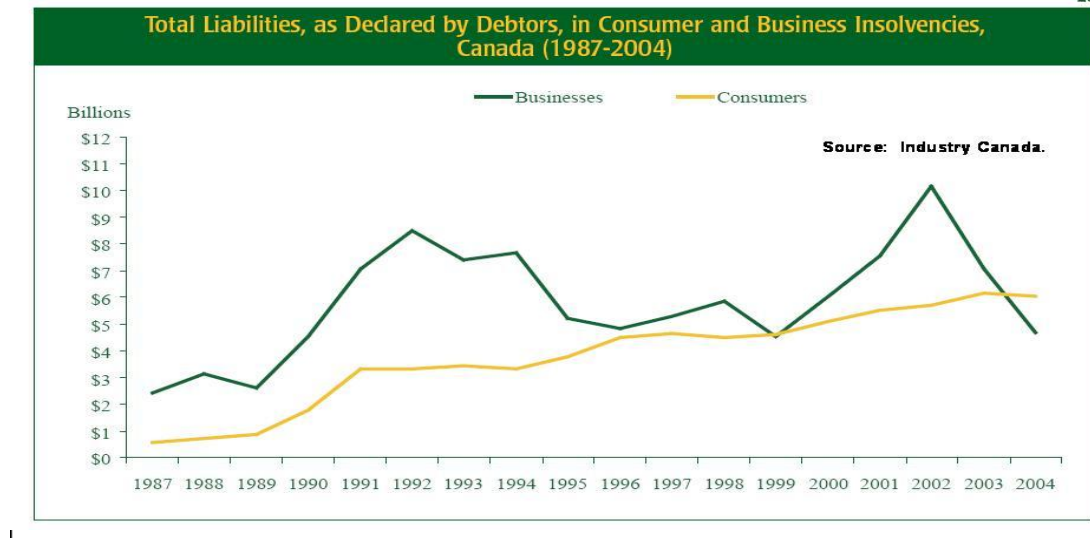


- Problem Description
- Predictor Construction
- Experimental Results
- Conclusions

Personal Bankruptcy Prediction

Why Personal Bankruptcy Prediction

- Big rise in personal bankruptcy leads to increasing losses to creditors

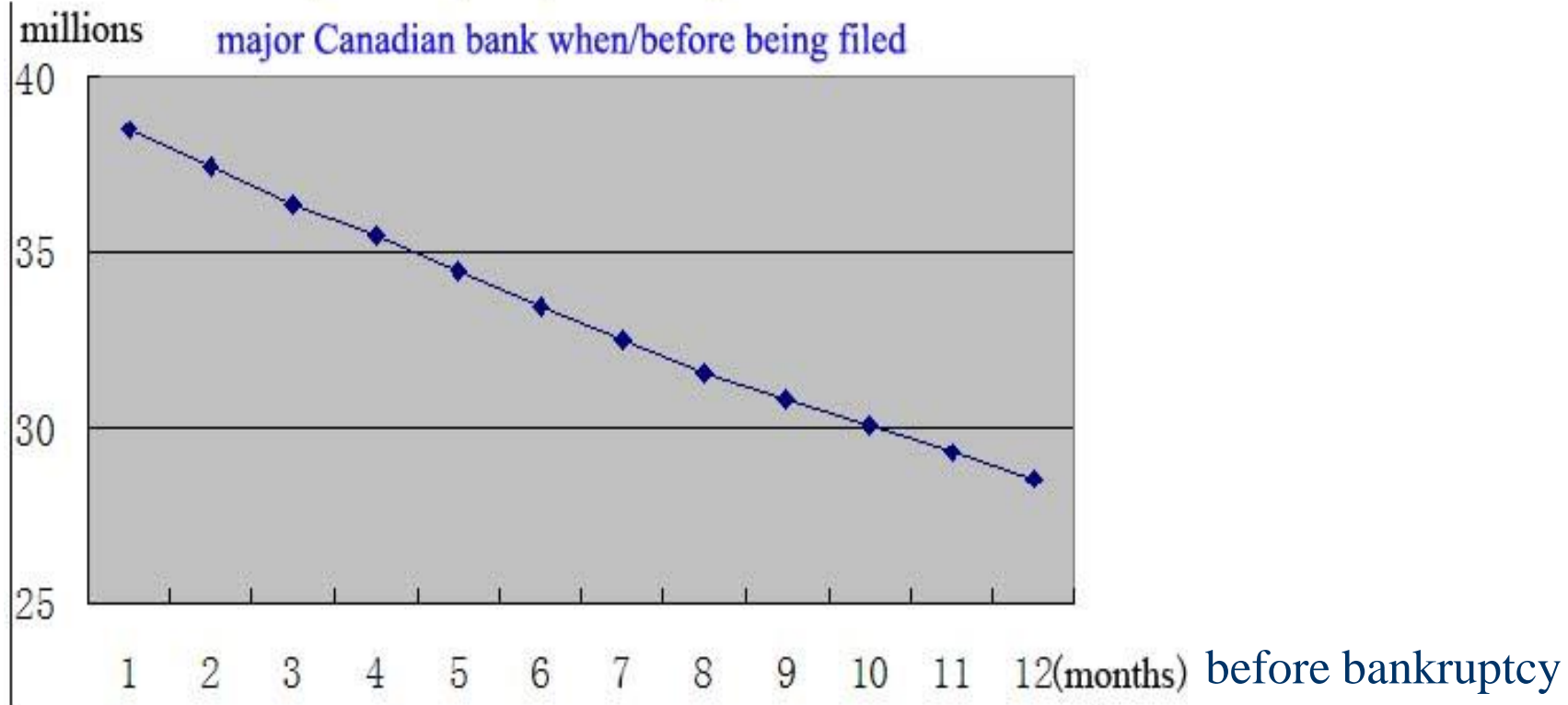


- Pre-granting strategies are not enough

The earlier, The better

Debts

Totally balance(debts) of bankrupt account of a major Canadian bank when/before being filed



late

early

Some Existing Models

- Credit Scoring (Equifax, TransUnion and Fair Isaac)
- Data Mining Model (Decision Tree, Neural Network etc.)

The pitfalls

- Difficult to select the predictors
- Difficult to aggregate the original sequential data
- Low interpretation ability

$$decision = f(X)$$



How to construct X ?

How to interpret f to creditor?

Our Model

- **Predict bankruptcy among credit card users**

87.4% of personal bankruptcy cases involve credit card debt

- **Construct predictor from sequence patterns**

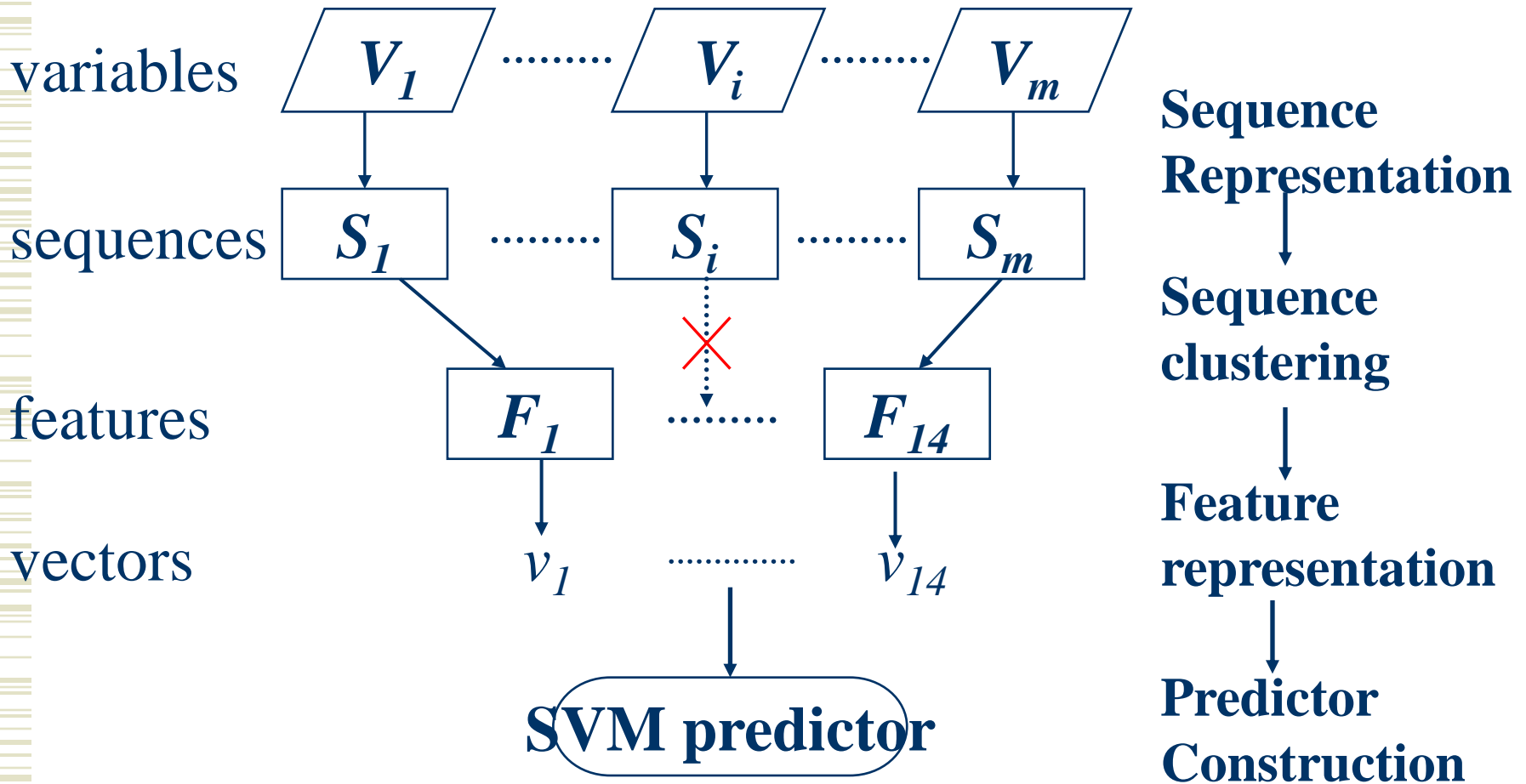
Discover sequences patterns as personal bankruptcy features

Purpose

- Extend existing method
- Improve prediction performance

Personal Bankruptcy Prediction System Using Sequence Mining Techniques

Framework of Our System

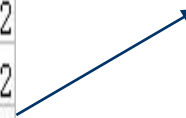


Sequence Representation of Client Behavior

- Sequences are constructed based on month time unit
- Binary sequence

ID	MONTH	BUREAU_SCORE	X_BALANCE	X_CASH_ADV	N_CASH_ADV
99753	200603	623	2754.88	66.59	2
99753	200602	663	2609.14	112.55	3
99753	200601	663	2577.28	51.53	2
99753	200512	663	2442.31	125.25	2
99753	200511	645	2387.67	37.74	2
99753	200510	645	2194.71	0	0
99753	200509	645	2230.75	46.3	3
99753	200508	671	2214.46	181.28	2
99753	200507	671	1972	0	0
99753	200506	671	2052.77	0	0
99753	200505	681	1988.19	0	0
99753	200504	681	2113.19	38.36	2

'111110110001'



A New Statistic Model for Sequence Cluster

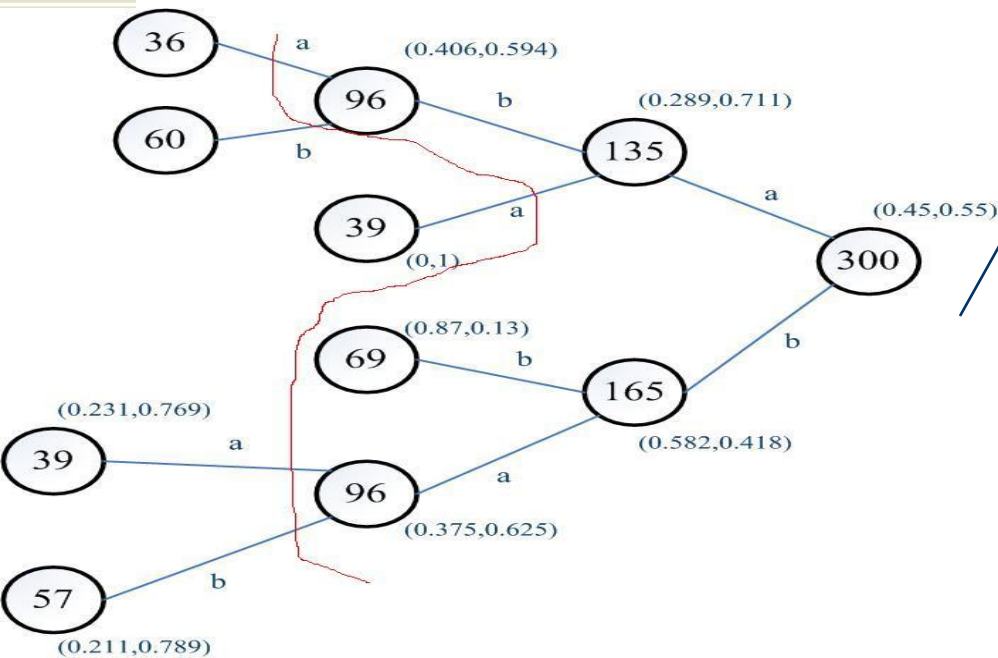
Sequence: $S = s_1 s_2 \cdots s_l$

Similarity between S and Δ :

The sequence cluster Δ :

$$sim_{\Delta}(S) = \prod_{i=1}^l \exp(w_i (P(s_i | s_1 \cdots s_{i-1}) - P(s_i)))$$

$$= \frac{\exp(\sum_{i=1}^l w_i \times P(s_i | s_1 \cdots s_{i-1}))}{\exp(\sum_{i=1}^l w_i \times P(s_i))}$$



Model-based *K*-means

Sequence Clustering results of one variable

cluster	#bankrupt	#non-bankrupt
1	1028	1450
2	638	106
3	96	184
4	120	168
5	118	92

**Bankruptcy
cluster**



Bankruptcy Feature Representation

Having 'Cash Advance' pattern:

2 consecutive cash advance followed by 1 with no cash advance

Month	Cash Advance	Cash Advance
200610	400	0
200609	0	100
200608	100	0
200607	200	50
200606	0	0
200605	0	0
200604	50	50
200603	0	100

Client 1

Client 2

Vector values:

Client 1: 300

Client 2: 0

Deal with Transaction Data

DATE_TRANSACTION	CODE_TRANSACTION	AMOUNT
10/3/2005 0:00	30	20
10/3/2005 0:00	36	2.5
10/5/2005 0:00	30	20
10/5/2005 0:00	30	20
10/5/2005 0:00	30	20
10/5/2005 0:00	36	2.5
10/5/2005 0:00	36	2.5
10/5/2005 0:00	36	2.5
10/7/2005 0:00	30	20
10/7/2005 0:00	30	40
10/7/2005 0:00	30	41.2

➤ **MCA**-> bankruptcy relevant transaction

➤ **Sequential Pattern Mining**-> bankruptcy transaction pattern

Experimental Results

The Data

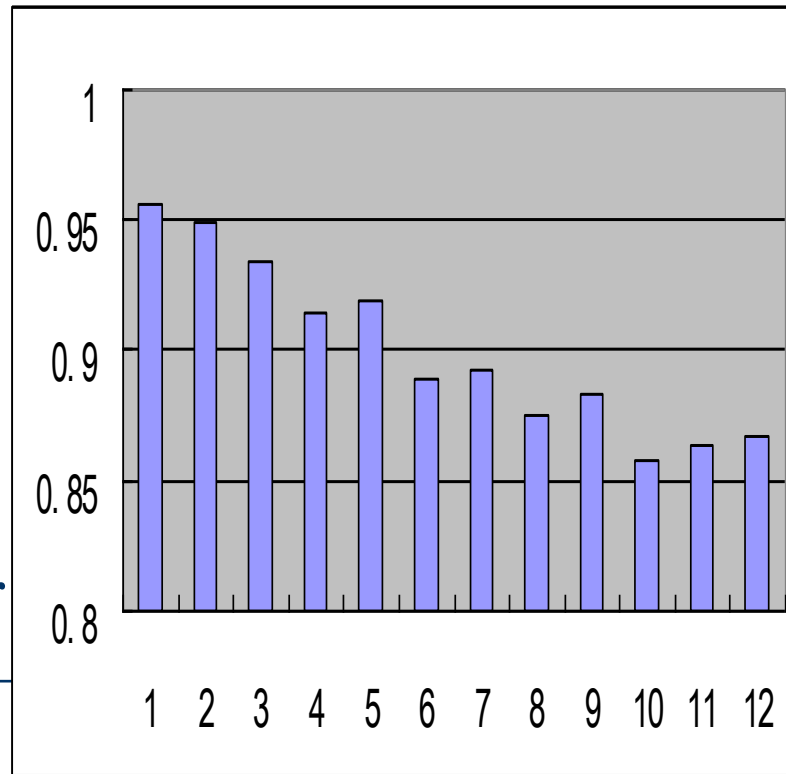
Data from a major Canadian bank

- Totally 7495 bad accounts +10,175 good accounts
- 2000 bad accounts +2000 good accounts Used for sequence analysis (bankruptcy features extraction); the remaining 5495 bad accounts and 8175 good accounts are used for evaluation

Time Line

Identification Ratio

one year



**more difficult
to make
longer period
prediction**

Historical
Period

Observation
point

Performance
Period

Comparison of Case Identification

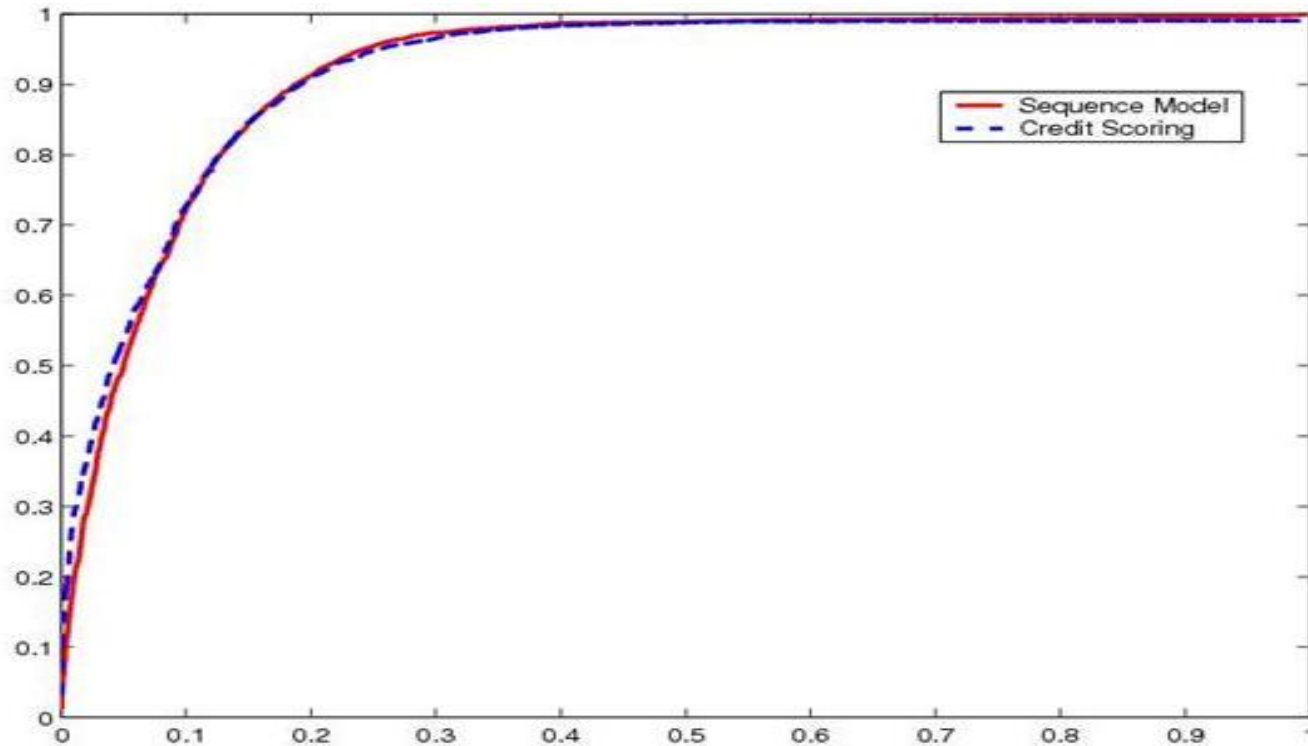
True \ False	10%	15%
Our system	70.92%	82.66%
Credit Scoring	72.68%	84.59%

False: fraction of # misidentified good account

True: fraction of # truly identified bad account

Comparison of Case Identification

True Identification



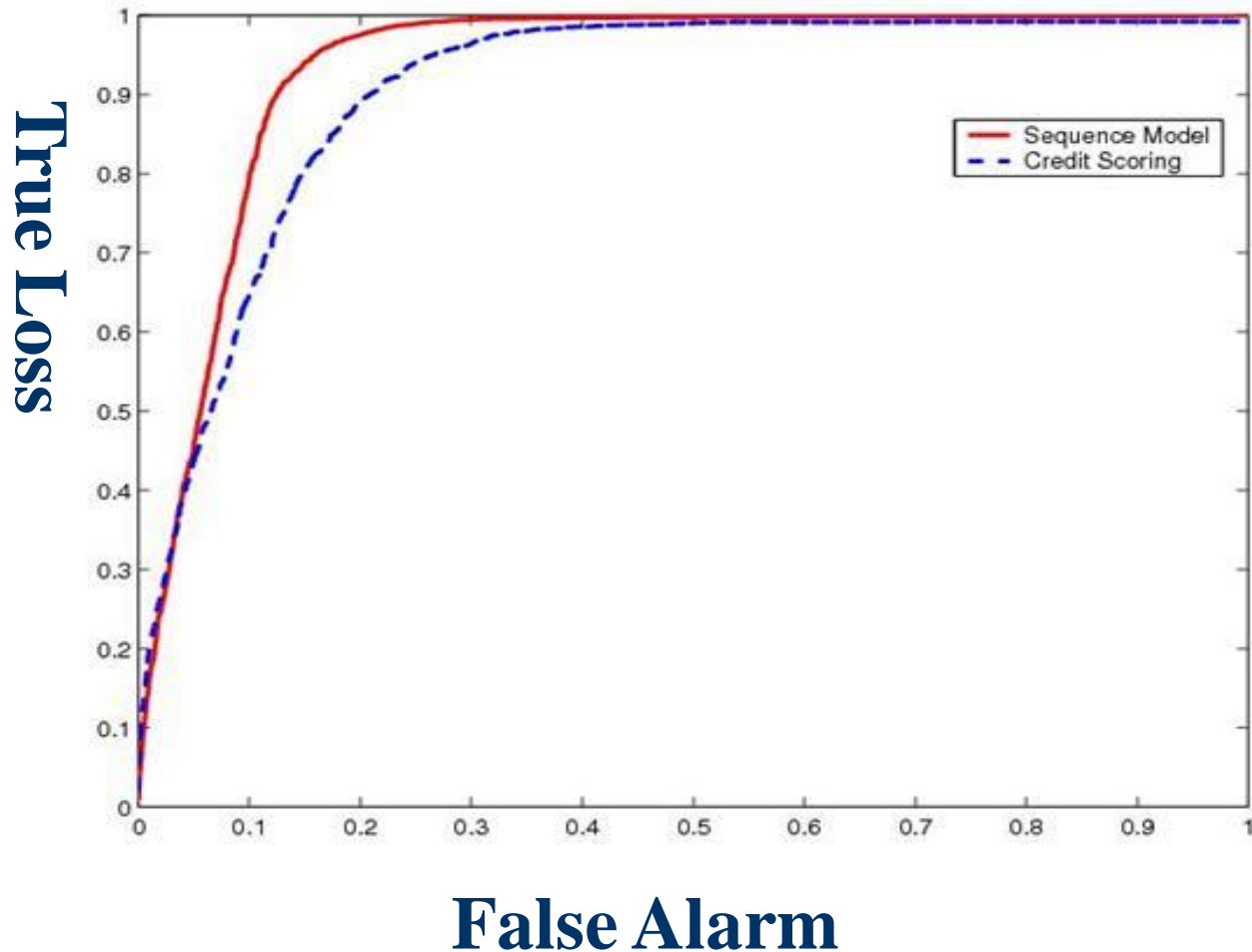
False Alarm

Comparison of Loss Prediction (observation point)

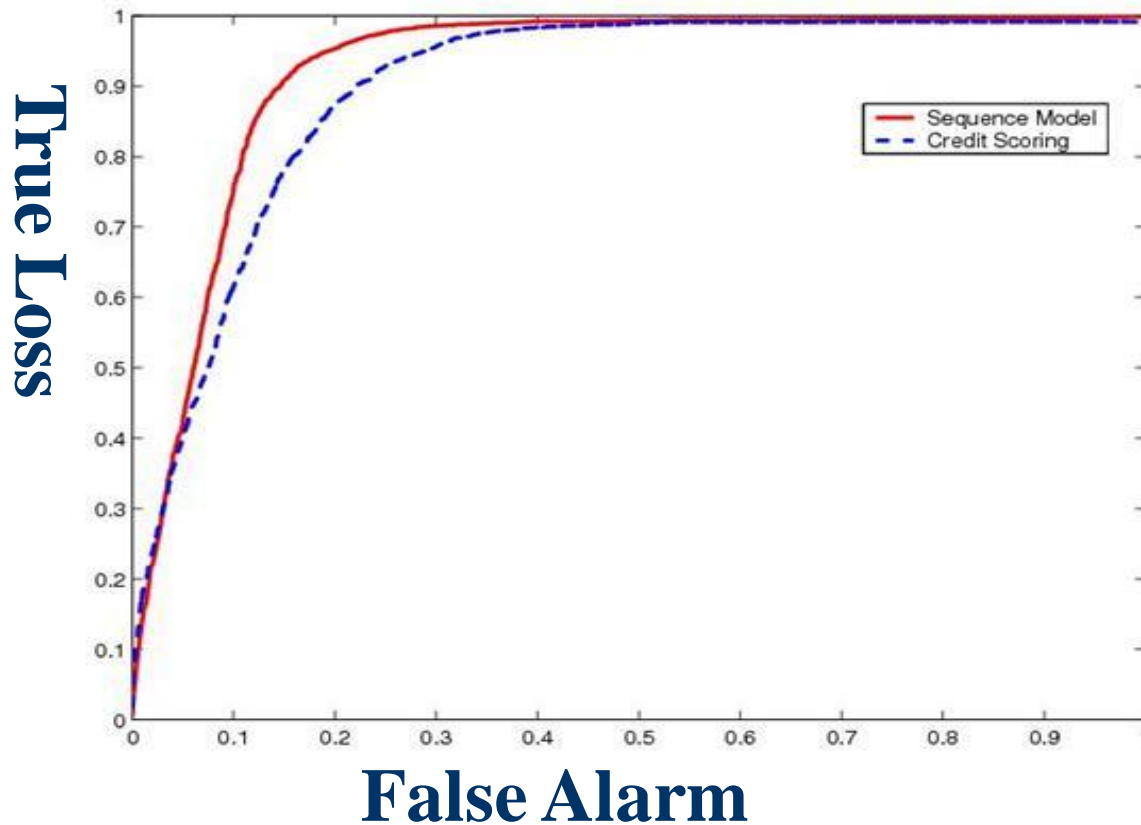
True Loss	False	10%	15%
	Our system	85.2%	92.75%
	Credit Scoring	64.65%	80.53%

True Loss: Fraction of balance of identified bad accounts on observation point

Comparison of Loss Prediction (observation point)



Comparison of Loss Prediction (bankruptcy point)



True Loss: Fraction of balance of identified bad accounts when declared bankruptcy

Conclusions

- Sequence patterns are very capable for identifying bad accounts
- These patterns can be obtained through *K*-means
- Our system outperforms credit scoring in case of loss prediction
- The system can be extended through mining transaction data