SIEMENS

Margin-Closed Sequential Pattern Mining

Dmitriy Fradkin, Fabian Moerchen

Siemens Corporate Research
Princeton, NJ

UP-2010

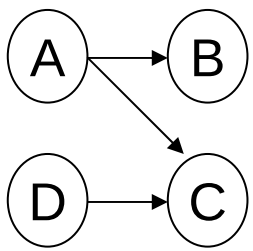
SIEMENS

Problem Statement

Objective: Given a database of item(set) sequences, find interesting/useful patterns. But not too many!

What are patterns: item(sets) with a partial order relation on them.

- Serial/Sequential: A->B->C
- Parallel: (A->B and D->C)
- Partial Order [Casas-Garriga 2005]: not equivalent to a combinations of serial-parallel episodes [Pei et al. 2006].



Classes of Patterns

Support of a pattern: number of occurrences / size of database.

Frequent patterns: support is above some pre-specified threshold.

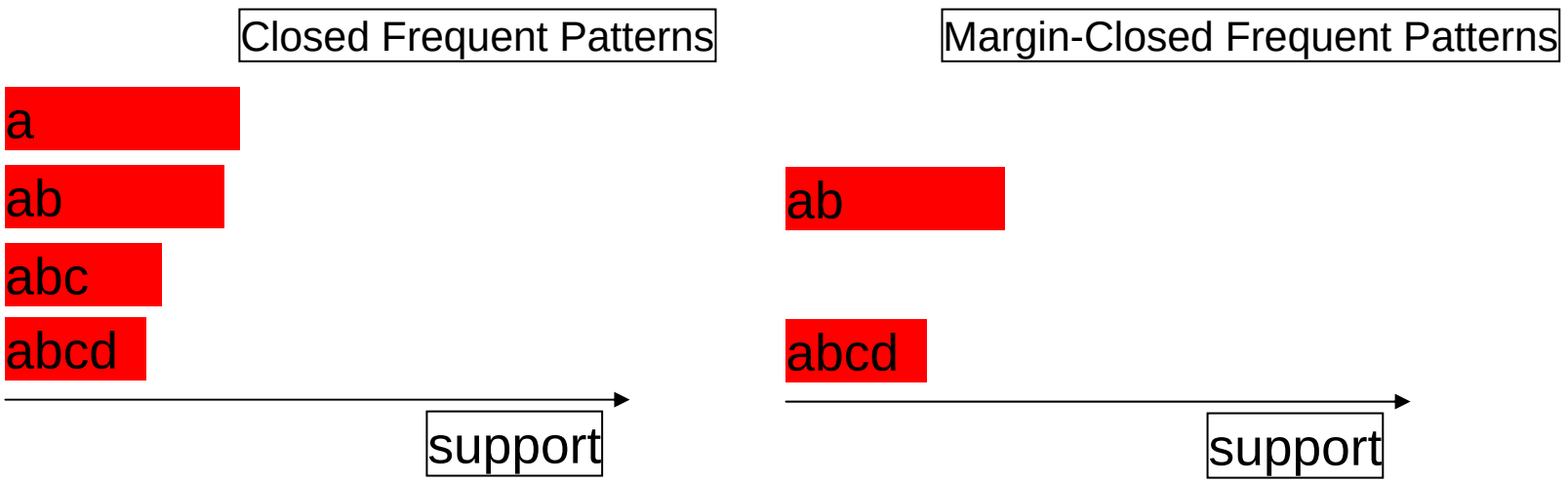
- Redundancy problem: Even if a sub-pattern has exactly same frequency as a super-pattern both are reported as frequent.

Closed frequent patterns: a pattern can't be extended without losing support.

- Still can produce a lot of patterns – sensitive to smallest difference in support.

Margin-Closed frequent patterns:

A pattern P with support S is margin-closed (with margin 0<alpha<=1) if there is no extension of P with support greater than (1-alpha)*S



SIEMENS

Related Work

Itemset Mining:

- Condensed representations [Calders et al 2006], constrained itemsets [Pei et. al 2001] and combinations
- Compression [Tatti & Vreeken 2008, Leeuwen et al. 2006].
- Non-derivable itemsets [Calders & Goethals, 2009] for frequency queries
- Margin-closed itemsets: used in [Moerchen 2006,Moerchen & Ultsch 2007, Cheng et.al 2006]. An efficient algorithm, extending DCI_Closed [Lucchese et al 2003], proposed in [Moerchen et.al, 2010].

Sequential mining:

- Compression of the mining result in a post-processing step [Cheng et. al. 2009, Wang 2008],
- Condensed representation to evaluate sequential association rules [Plantevit & Crémilleux, 2009],
- Approximate patterns [Zhu 2007] under the Hamming distance.

Our approach:

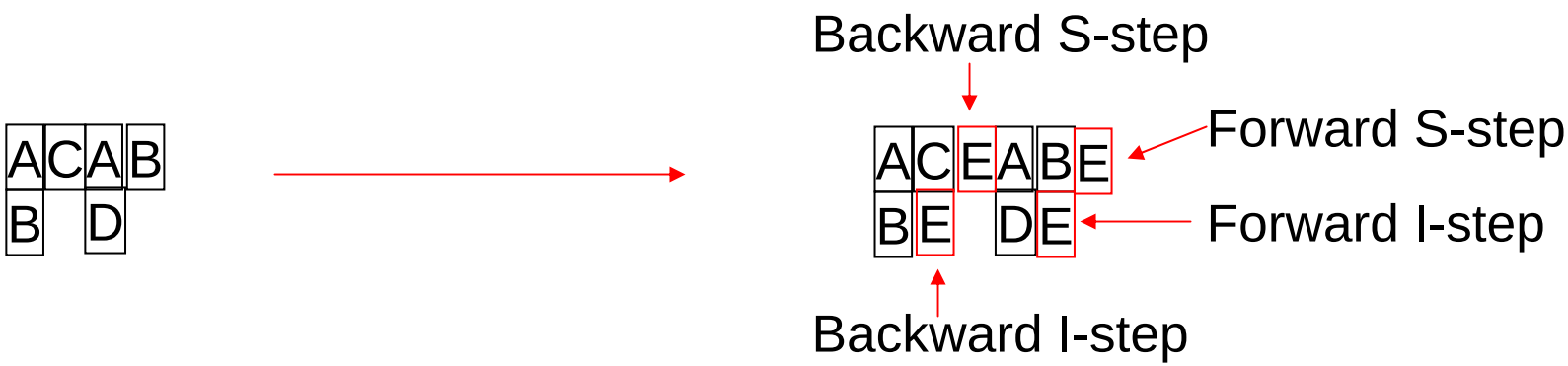
- Favors longer patterns, whereas condensed representation focus at reconstruction of frequencies for patterns not reported or compression ratio of the complete pattern set.
- Only observed patterns, with exact frequencies, are reported, unlike whereas approximate patterns.
- The pruning is integrated in the mining algorithm whereas compression is a post-processing of the results after mining.

SIEMENS

BIDE: Pattern Extensions

Many algorithms for mining closed sequential patterns – we extend BIDE - BIDirectional Extension checking [Wang & Han 2004]

Pattern Extension: There are 4 ways to extend a sequential pattern:



Theorem [Wang/Han 2004]: A pattern that can't be extended with any of the above steps without losing support is closed.

BIDE: Main Algorithm

Find all frequent items in the projected database

If any forward extension or backward extension items have the same support as pattern, then pattern is not closed.

Extend pattern with frequent items. If the extended pattern will be reached in a different way (checked by backscan), it can be pruned currently. Otherwise, recursively evaluate the extended pattern.

Algorithm 1 BIDE Algorithm

Require: Sequential Pattern $P = \{p_i\}$, Projected Database $D|P$, minimum support μ

1: F - set of frequent closed patterns
2: $l = |P|$
3: $Ls = sStepFrequentItems(P, D|P, \mu)$;
4: $Li = iStepFrequentItems(P, D|P, \mu)$;
5: **if** $!(frequencyCheck(Ls, P) \parallel frequencyCheck(Li, P))$ **then**
6: **if** $backscan(P', D', true)$ **then**
7: $F = F \cup P$
8: **end if**
9: **end if**
10: **for** itemset $p \in Ls$ **do**
11: $P' = p_1, .., p_l, p$
12: **if** $backscan(P', D|P', false)$ **then**
13: $F = F \cup bide(P', D', \mu)$;
14: **end if**
15: **end for**
16: **for** itemset $p \in Li$ **do**
17: $P' = p_1, .., p_{l-1}, p_l \cup p$
18: **if** $backscan(P', D|P', false)$ **then**
19: $F = F \cup bide(P', D', \mu)$;
20: **end if**
21: **end for**
22: **return** F

SIEMENS

BIDE-Margin: Enforcing Margin-closedness

We now describe the changes required to enforce margin-closedness in BIDE leading to the BIDE-Margin algorithm.

1. When Forward Expansion (frequencyCheck function) is considered, rather than checking if there are items with the same support as the current pattern, instead **check for presence of items that are within margin** of the pattern's support.
2. When checking backward closure (first call to backscan function), **look if there are any items that are margin-close to the pattern** i.e. if there is an item with frequency above and within margin of the support of P, we know that P is not margin-closed.

Note: When checking if a pattern will be reached via an alternative path, the above check should not be performed – we cannot disregard the recursion branches going from the current pattern unless there is a backward extension with **exactly** the same support.

Computational Complexity:

BIDE-Margin generates all frequent sequential patterns just as BIDE does, in exactly the same fashion.

However, since it searches for margin closed patterns, it will need to call backscan function less frequently.

SIEMENS

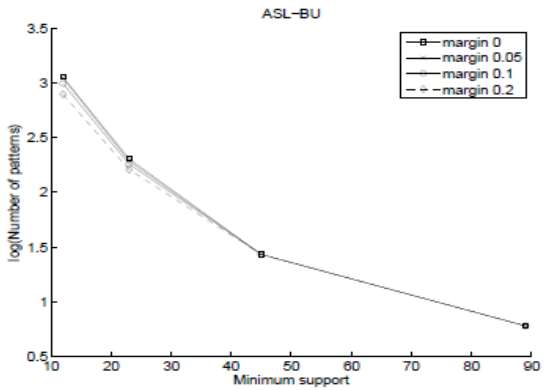
Data

Data	Intervals	Labels	Sequences	Classes
ASL-BU	18250	154	441	7
Auslan2	900	12	200	10
Blocks	1207	8	210	8
Context	12916	54	240	5
Pioneer	4883	92	160	3
Skating	18953	41	530	6/7

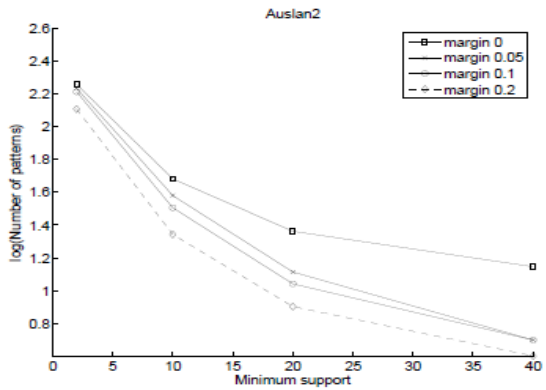
While technically databases of intervals, these can be interpreted as sequential databases by treating start and end boundaries of an interval as separate events [Wu & Chen 2007, Fradkin & Moerchen 2010].

SIEMENS

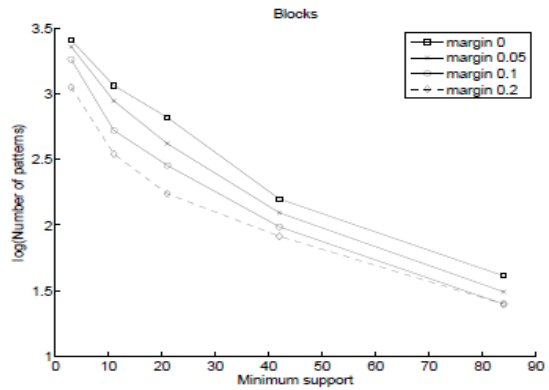
Experiment: Numerosity



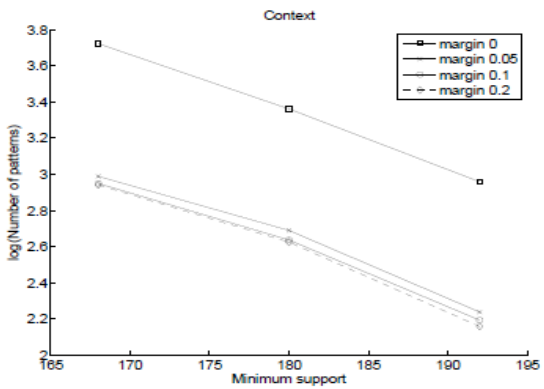
(a) ASL-BU



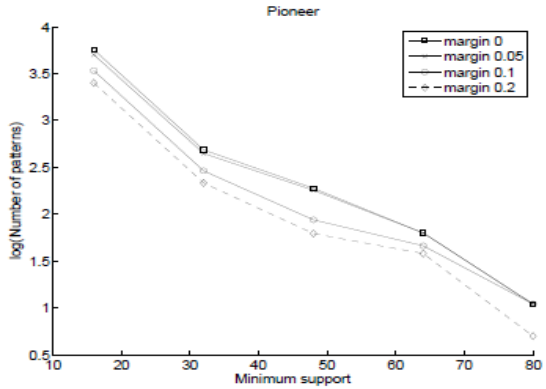
(b) Auslan2



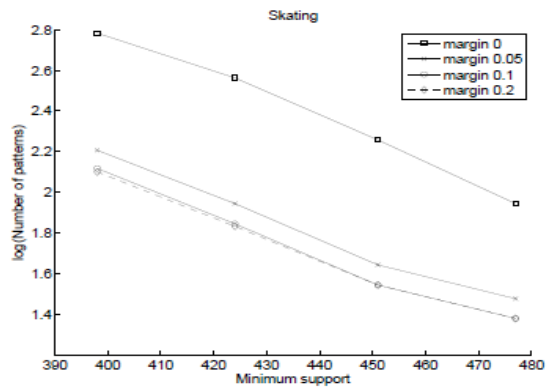
(c) Blocks



(d) Context



(e) Pioneer



(f) Skating

SIEMENS

Predictiveness Experiments

- **Representation:** Patterns are used as binary features.

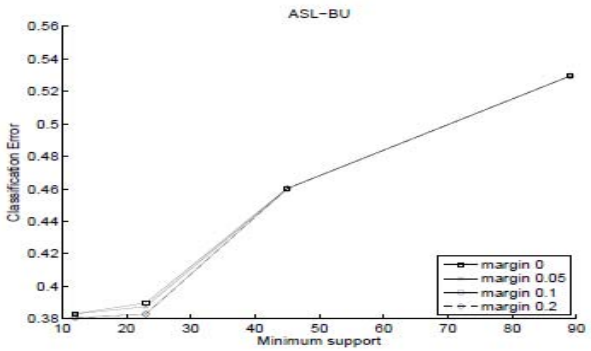
- Spider Toolbox for Matlab:
  - 10-fold cross-validation
    - **Support Vector Machines**: C=2^k, k = -10,…,10 – best result reported
    - **J48**, default settings
    - **Random Forests**

Results:

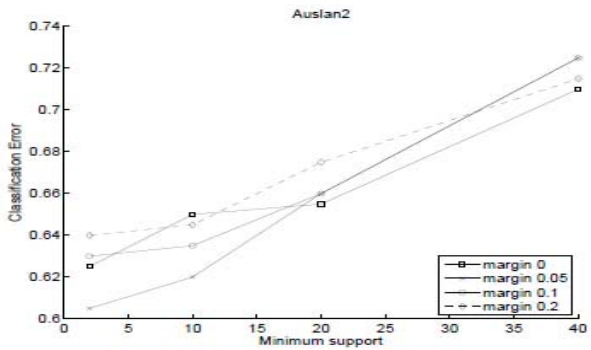
- using margin 0.05 or 0.1 barely affects the classification error rate.
- margin of 0.2 does lead to noticeably worse results on Pioneer dataset, and on Auslan2 with support 20, but not on the other datasets.
- The differences in performance tend to become smaller as support increases and the number of patterns decreases.

SIEMENS

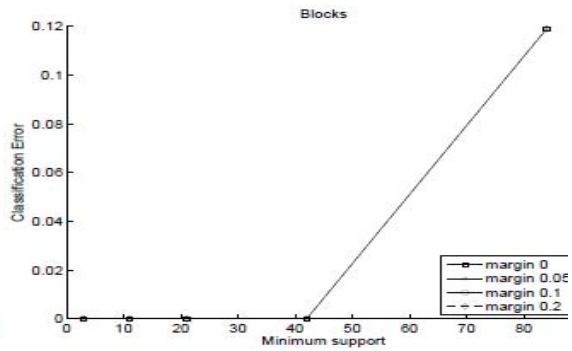
Experiment: Predictiveness with SVM



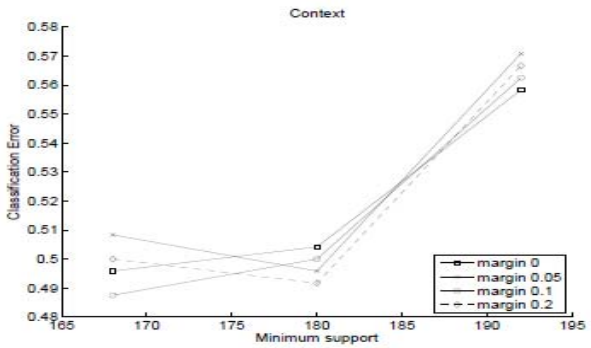
(a) ASL-BU



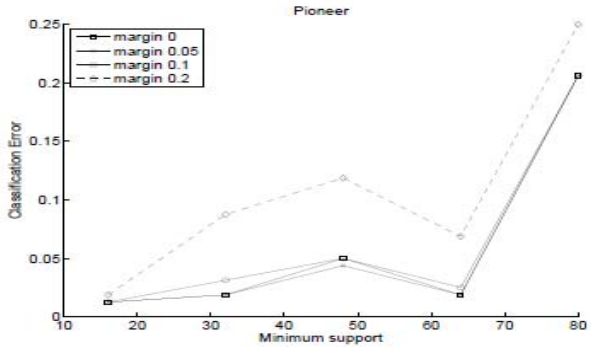
(b) Auslan2



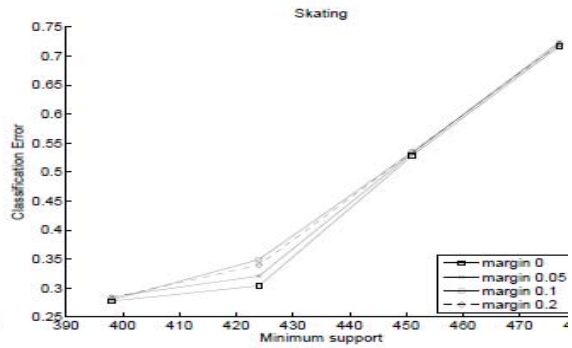
(c) Blocks



(d) Context



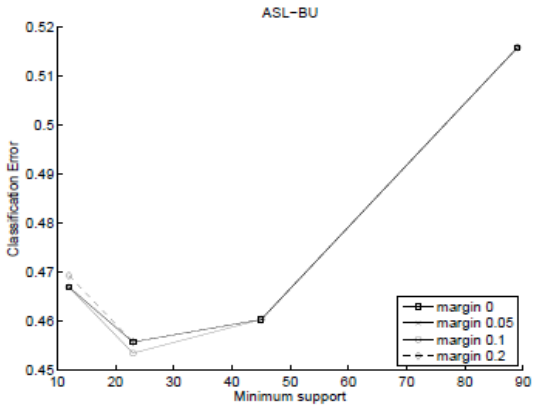
(e) Pioneer



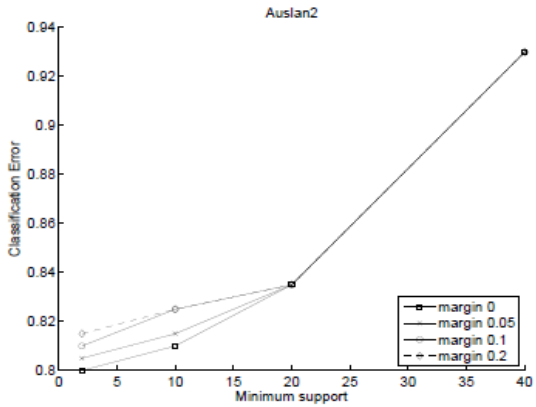
(f) Skating

SIEMENS

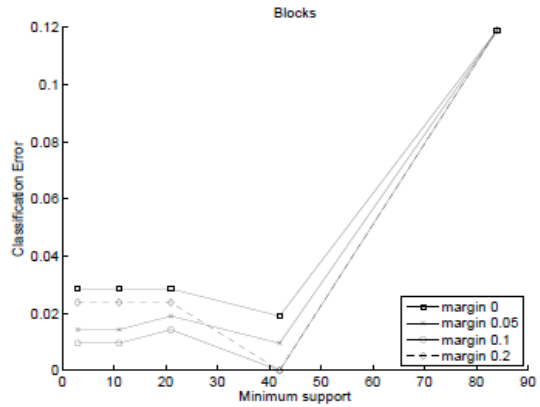
Experiment: Predictiveness with J48



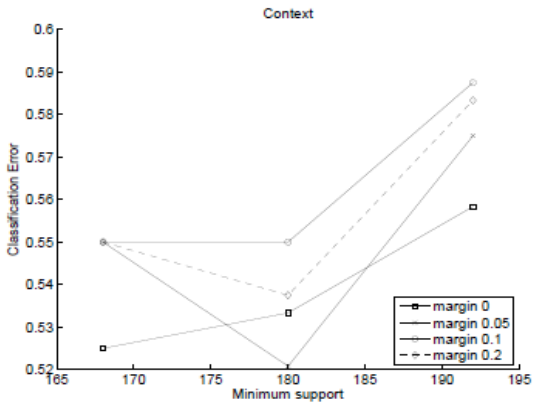
(a) ASL-BU



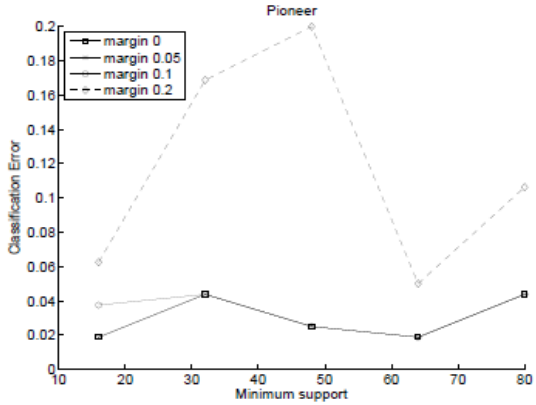
(b) Auslan2



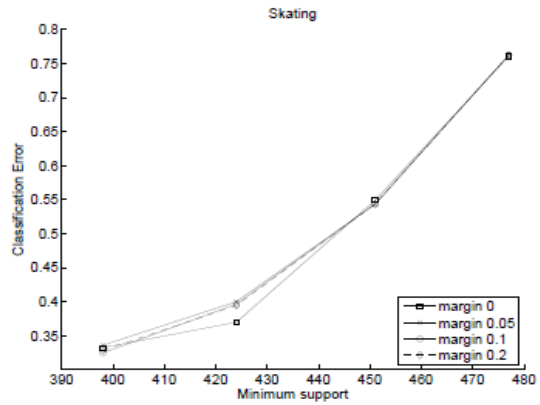
(c) Blocks



(d) Context



(e) Pioneer



(f) Skating

SIEMENS

Conclusion

- Presented a new constraint for reducing the output of sequential pattern mining
- Presented an efficient algorithm for mining such patterns:
  - exact patterns with exact frequency are reported!
- Demonstrated that the number of margin-closed patterns can be a lot smaller than that of closed patterns
- The resulting patterns are just as useful, as evidenced by performance of classifiers built using these patterns.