

Approximate Frequent Itemsets

Selim Mimaroglu

Find this presentation at

<http://www.cs.umb.edu/~smimarog>

References

1. “Mining Approximate Frequent Itemsets from Noisy Data” (ICDM 05) by Jinze Liu, Susan Paulsen, Wei Wang, Andrew Nobel, Jan Prins
2. “Efficient Discovery of Error-Tolerant Frequent Itemsets in High Dimensions” by Cheng Yang, Usama Fayyad, Paul S. Bradley

[1] is the main reference of this presentation.

Interesting Observations

- Approximate Itemsets in the sense that some itemsets which should have been recorded as “1” might have been recorded as “0”
- Above is the definition of *noise* in [1] (but not vice versa)
- Both [1] and [2] generate frequent itemsets which are *missed* by the exact Apriori
- Therefore [1], and [2] require more time compared with exact Apriori

Error in row, ϵ_r , is permitted

- $\epsilon_r \in [0, 1]$
- for each $t \in T$ the fraction of items in I that appear in t is at least $(1 - \epsilon_r)$
- $I = \{a, b, c, d, e\}$ is a AFI (Approximate Frequent Itemset) with minsup = 100% and $\epsilon_r = 0.2$

	a	b	c	d	e	$\epsilon_r\%$
1	1	1	1	1	0	20
2	1	1	1	1	0	20
3	1	1	1	1	0	20
4	1	1	1	1	0	20
5	1	1	1	1	0	20

Only ϵ_r isn't enough

- There may be some *free riders* (Main improvement of [1] over [2])
- e of $I=\{a,b,c,d,e\}$ with $minsup = 100\%$ and $\epsilon_r = 0.2$ is a free rider!

	a	b	c	d	e	$\epsilon_r\%$
1	1	1	1	1	0	20
2	1	1	1	1	0	20
3	1	1	1	1	0	20
4	1	1	1	1	0	20
5	1	1	1	1	0	20

Error in column, ϵ_c , is permitted

- $\epsilon_c \in [0, 1]$
- for each $i \in I$ the fraction of transactions in T that appear in each item i is at least $(1 - \epsilon_c)$
- $I = \{a, b, c, d, e\}$ is a AFI (Approximate Frequent Itemset) with minsup = 100% and $\epsilon_c = 0.2$

	a	b	c	d	e
1	1	1	1	1	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	0	0	0	0	0

$\epsilon_c \%$ 20 20 20 20 20

Only ϵ_c isn't enough

- AFI $I=\{a,b,c,d,e\}$ with $minsup = 100\%$ and $\epsilon_c = 0.2$, transaction 5 is a free rider!

	a	b	c	d	e
1	1	1	1	1	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	0	0	0	0	0
$\epsilon_c\%$	20	20	20	20	20

Both ϵ_r and ϵ_c is needed

- Definition 1.1 [1]

Let D be a (binary) data matrix, and let $\epsilon_r, \epsilon_c \in [0, 1]$. An itemset $I \subseteq I_0$ is an *AFI*, if there exists a set of transactions $T \subseteq T_0$ with $|T| \geq \text{minsup} \cdot |T_0|$ such that following two conditions hold

1. for each $t \in T$ the fraction of items in I that appear in t is at least $(1 - \epsilon_r)$
2. for each $i \in I$, the fraction of transactions in T that appear in each item i is at least $(1 - \epsilon_c)$

Example

- Example 1.1 [1] (This example refers to Table 1, not Fig 1(A))
- $minsup=0.5$, $\varepsilon_r=1/3$, and $\varepsilon_c=1/3$
- A maximal *AFI* contained in D is $I = \{a,b,c\}$
($T=\{1,2,3,4,5\}$)

	a	b	c	d
1	1	1	1	0
2	1	1	0	0
3	1	0	1	0
4	0	1	1	0
5	1	1	1	1
6	0	0	0	1
7	0	1	0	1
8	1	0	0	0

Conclusion

- This is useful, but the definition of noise doesn't fit to practical applications. Noise is defined as $1 \Rightarrow 0$, but it may also happen that $0 \Rightarrow 1$. Approach taken is *optimistic*.
- A property of sets P is *antimonotone* if a set S has a property P , all its subsets has the same property P . This feature is used in Apriori, but since ε_r , and ε_c is introduced it doesn't hold anymore, therefore discovering *AFIs* is more time consuming than Apriori.