Items2Data: Generating Synthetic Boolean Datasets from Itemsets

Ian Wong Gillian Dobbie Yun Sing Koh

School of Computer Science
University of Auckland

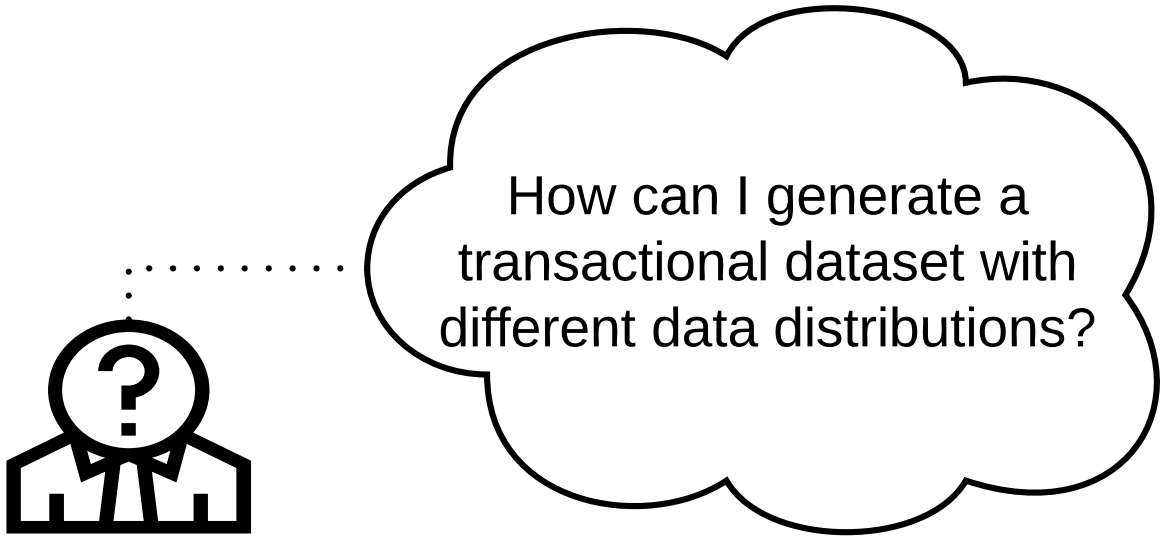
iwon015@aucklanduni.ac.nz

February 17, 2019

Acknowledgement: CORE Student Travel Award

Overview

Synthetic Data is Important in Machine Learning



Independent

a	b	c

Conditionally Independent

a	b	c

Conditionally Dependent

a	b	c

Typical Multivariate Data Generation

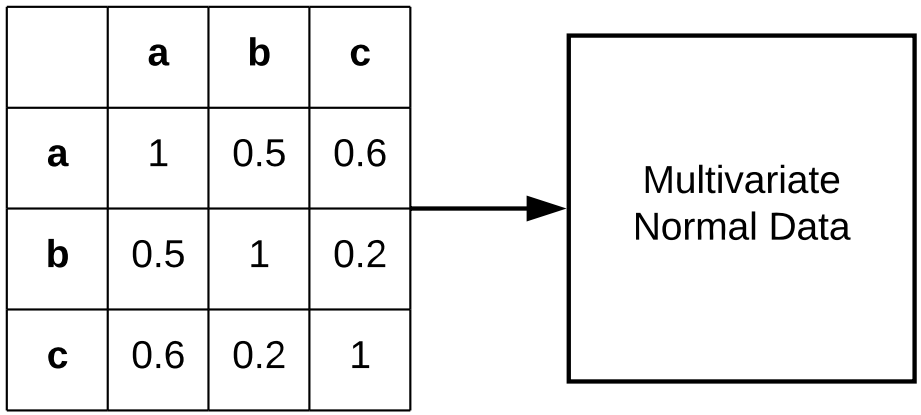


Figure: Define a correlation matrix and generate data which satisfies it

However...

- Correlation matrices are limited to pairwise correlations
- Popular matrix factorisation techniques such as the Cholesky Decomposition[1] do not generate boolean data

[1] L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 1997. ISBN: 9780898713619. URL: https://books.google.co.nz/books?id=bj-Lu6zjWbEC.

Boolean Data is Very Common

Boolean / Binary

is_X	is_X
True	1
False	0

Categorical

Gender	Gender_M	Gender_F
M	1	0
F	0	1

Transactional

TID	**Items**
1	**a, b**
2	**b, c**
3	**c**

a	b	c
1	1	0
0	1	1
0	0	1

Alternative to Correlations: Itemset Representations

Instead of using correlation matrices to design boolean data, we might be able to use itemsets.

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

itemset	count	support
a	4	0.500
b	4	0.500
c	3	0.375
ab	2	0.250
ac	2	0.250
bc	2	0.250
abc	1	0.125

Figure: A transaction database with items **a**, **b** and **c** alongside the corresponding itemset representation.

Advantages of Itemsets

- Easy to understand
- Correlation between items **a** and **b** is defined by the support of **a**, **b** and **ab**
- Complex relationships can be defined with longer itemsets e.g. **abc** defines the joint distribution between **a**, **b** and **c**

How to Generate Data from Itemsets?

Inverse Frequent Itemset Mining (IFM)[2]:

- Let D be a database of transactions and S be a set of itemsets with corresponding support
- Task: Generate a database D that satisfies the supports of itemsets in S
- Proven to be NP-Hard
- Efficient approaches sacrifice accuracy in satisfying S for improved efficiency, which limits practical application[3]

[2] Taneli Mielikainen. "On Inverse Frequent Set Mining". In: *Proc. of the 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining*. Citeseer. 2003, pp. 18–23.

[3] Antonella Guzzo, Domenico Saccà, and Edoardo Serra. "An Effective Approach to Inverse Frequent Set Mining". In: *Ninth IEEE International Conference on Data Mining, 2009. ICDM'09.* IEEE. 2009, pp. 806–811.

Contributions

- An algorithm, Items2Data, that exactly satisfies itemset supports in $O(n^2)$ time when the itemsets in S are *globally closed*
- In defining the algorithm we also introduce concepts of marginal support and global closure

Items2Data: What is Marginal Support?

The marginal support of an itemset is the itemset's support minus the marginal support of all its supersets.
It represents unique information that is not explained by its supersets.

Itemset	Support	Marginal Support	Marginal Formula
abc	0.125	0.125	s(abc)
bc	0.25	0.125	s(bc) - m(abc)
ac	0.25	0.125	s(ac) - m(abc)
ab	0.25	0.125	s(ab) - m(abc)
c	0.375	0	s(c) - m(ac) - m(bc) - m(abc)
b	0.5	0.125	s(b) - m(ab) - m(bc) - m(abc)
a	0.5	0.125	s(a) - m(ab) - m(ac) - m(abc)
{}	1	0.25	s({}) - m(abc) - m(bc) - m(ac) - m(ab) - m(c) - m(b) - m(a)

Figure: $s(I)$ = support of itemset I, $m(I)$ = marginal support of itemset I

Items2Data: What is Global Closure?

A set of itemsets is *globally closed* if sum(Marginal Support) ≤ 1

Itemset	Support	Marginal Support	Marginal Formula
abc	0.125	0.125	s(abc)
bc	0.25	0.125	s(bc) - m(abc)
ac	0.25	0.125	s(ac) - m(abc)
ab	0.25	0.125	s(ab) - m(abc)
c	0.375	0	s(c) - m(ac) - m(bc) - m(abc)
b	0.5	0.125	s(b) - m(ab) - m(bc) - m(abc)
a	0.5	0.125	s(a) - m(ab) - m(ac) - m(abc)
		0.750	

Figure: A set of globally closed itemsets

Items2Data: What is Global Closure?

A set of itemsets is **not** *globally closed* if sum(Marginal Support) > 1

Itemset	Support	Marginal Support	Marginal Formula
c	0.375	0.375	s(c)
b	0.5	0.5	s(b)
a	0.5	0.5	s(a)
		1.375	

Figure: A set of **not** globally closed itemsets

Items2Data: Why does Global Closure Matter?

Global Closure describes when IFM is Tractable

Itemset	Support	Marginal Support	Marginal Formula
abc	0.125	0.125	s(abc)
bc	0.250	0.125	s(bc) - m(abc)
ac	0.250	0.125	s(ac) - m(abc)
ab	0.250	0.125	s(ab) - m(abc)
c	0.375	0.000	s(c) - m(ac) - m(bc) - m(abc)
b	0.500	0.125	s(b) - m(ab) - m(bc) - m(abc)
a	0.500	0.125	s(a) - m(ab) - m(ac) - m(abc)
{}	1.000	0.250	s({}) - m(abc) - m(bc) - m(ac) - m(ab) - m(c) - m(b) - m(a)

a	b	c	|D| = 8
1	**1**	**1**	+1 **abc**
0	**1**	**1**	+1 **bc**
1	0	**1**	+1 **ac**
1	**1**	0	+1 **ab**
0	**1**	0	+1 **b**
1	0	0	+1 **a**
0	0	0	+2 **{}**
0	0	0	

Figure: When a set of itemsets is *globally closed*, the marginal support corresponds to a proportional number of rows in a satisfying dataset

Items2Data: Why does Global Closure Matter?

Itemset	Support	Marginal Support	Marginal Formula
c	0.375	0.375	s(c)
b	0.5	0.5	s(b)
a	0.5	0.5	s(a)

a	b	c	|D| = 8
0	0	**1**	+3 **c**
0	0	**1**	
0	0	**1**	
0	**1**	0	+4 **b**
0	**1**	0	
0	**1**	0	
0	**1**	0	
1	0	0	+ 4 **a?**

Figure: When itemsets are not globally closed how do we generate a database that satisfies support constraints? (Intractable)

Items2Data: Algorithm

1. Order a set of itemsets by longest itemset length to shortest resulting in S
2. Calculate the marginal support for each itemset in S starting from the longest resulting in a set of marginal supports M
3. If the sum $(M) \leq 1$ then S is globally closed
4. If globally closed then continue, else terminate
5. For each marginal support in M, add the equivalent number of rows to a database D where $|D|$ is user defined

Results: 100% Reconstruction Accuracy

Table: Dataset Characteristics and Reconstruction Times

Properties	accidents	bms1	bms2	chess	connect4	pumsb
#Transactions	340,183	59,602	77,512	3,196	67,557	49,046
#Unique Itemsets	**339,898**	18,473	48,684	3,196	67,557	48,474
%Unique Itemsets	0.9992	0.3099	0.6281	1.000	1.000	0.9883
#Items	468	497	3,340	75	129	2,113
Average Itemset Length	33.81	2.51	4.62	37.00	43.00	74.00
Support2Marginal Time	NA	22s	183s	1s	586s	263s
Marginal2Data Time	NA	25s	70s	1s	4s	24s
Reconstruction Accuracy	NA	100%	100%	100%	100%	100%

Results: Time Complexity

1. Order a set of itemsets by longest itemset length to shortest resulting in S $O(n \log n)$
2. Calculate the marginal support for each itemset in S starting from the longest resulting in a set of marginal supports M $\frac{n(n-1)}{2} \rightarrow O(n^2)$
3. If the sum $(M) \leq 1$ then S is globally closed $O(n)$
4. If globally closed then continue, else terminate $O(1)$
5. For each marginal support in M, add the equivalent number of rows to a database D where $|D|$ is user defined. $O(n)$

Complexity: $O(n^2)$ where $n = |S|$

Results: Execution Time

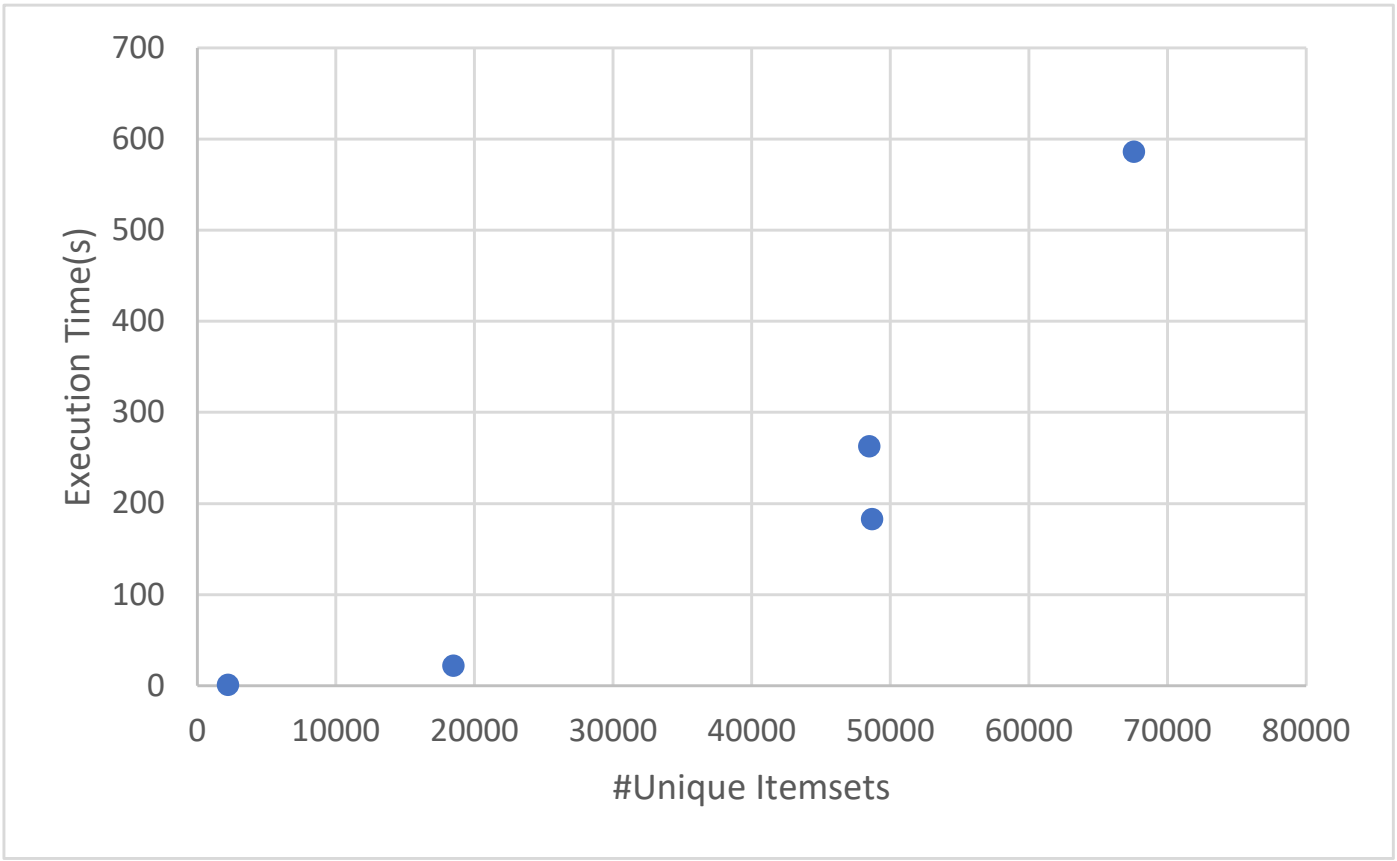


Figure: Execution time vs Unique Itemsets

Future Work

- Global closure is a powerful condition which describes when IFM tractable, however not all sets of itemsets are globally closed
- Introduce efficient itemset repair mechanisms to satisfy global closure
- Efficiently solve the IFM problem

Conclusion

- There are efficient techniques for generating synthetic real valued data
- Generating synthetic boolean data is hard
- Items2Data efficiently and exactly generates boolean data that satisfies a set of itemsets in quadratic time if itemsets are *globally closed*
- Demonstrated its speed and practical applicability on reconstructing real world data

Conclusion

- There are efficient techniques for generating synthetic real valued data
- Generating synthetic boolean data is hard
- Items2Data efficiently and exactly generates boolean data that satisfies a set of itemsets in quadratic time if itemsets are *globally closed*
- Demonstrated its speed and practical applicability on reconstructing real world data

Thank you!
Any Questions?
iwon015@aucklanduni.ac.nz