Synthetic Datasets

Rong Huang, Rada Chirkova, Yahya Fathi

1 Introduction

Datasets may be generated by algorithms in the purpose of testing the performance of database management systems. In this report, we will define symmetric synthetic dataset and two types of non-symmetric synthetic datasets that has some special structures and properties. The rest of the report is organized as follows. In Section 2, we introduce symmetric synthetic dataset, its structure and the properties of the associated views. In Section 3, we discuss two types of non-symmetric synthetic datasets by generalization and modification of the symmetric synthetic dataset. We end the report with some concluding remarks in Section 4.

2 Symmetric synthetic dataset

2.1 Construction of dataset

Define a dataset D as follows. The dataset D has K attributes. Each attribute takes m different values. We denote the dataset as $D(K, m)$. The master table contains all the possible entries by taking different values over K attributes. Hence, the number of rows in the master table is m^K. Under this definition, the dataset with a symmetric structure, is thus called *symmetric synthetic dataset*.

For example, assume dataset $D(3, 4)$ has 3 attributes A, B and C. Each attribute takes four values 0, 1, 2 or 3. The master table has 64 ($= 4^3$) entries. And the master table is shown in Table 1.

2.2 Views of symmetric synthetic dataset

A view, as a common type of derived data, is a virtual table consists of the result set of a query. In this context, we denote a view by its associated attributes. In a synthetic dataset $D(K, m)$, there are 2^K different associated views. Evaluate the size of a view by its number of rows. Then, the size of a k-attribute view is m^k. In a symmetric synthetic dataset, if view V_1 has the same number of attributes as view

V_2, then they have identical size. And if view V_1 is a decedent of view V_2, then V_2 is at least m times the size of V_1.

For example, $D(3, 4)$ has 8 (=2^3) views. Each view which contains only one attribute (view $\{A\}$, $\{B\}$ or $\{C\}$) has 4 (= 4^1) rows, shown in Table 2. And each view which contains two attributes (view $\{A, B\}$, $\{B, C\}$ or $\{A, C\}$) has 16 (= 4^2) rows, shown in Table 3. And the raw-data view (view $\{A, B, C\}$) has 64 (= 4^3) rows. View $\{A, B\}$ is four times the size of view $\{A\}$.

The size of views in $D(4, 7)$ and $D(5, 9)$ are shown in Table 4 and Table 5, respectively.

3 Non-symmetric synthetic dataset

In this section, we define two types of non-symmetric synthetic dataset based on the the symmetric synthetic dataset obtained in the previous section.

3.1 Type I non-symmetric synthetic dataset

3.1.1 Construction

We consider a generalization of the symmetric synthetic datasets by changing the condition that all the attributes take identical number of values. Define a dataset D as follows. The dataset D has K attributes, denoted as $a_1, a_2, \ldots, a_K$. Attribute a_k takes m_k different values, for $k = 1, 2, \ldots, K$. We denote the dataset as $D(K; m_1, \ldots, m_K)$. The master table contains all the possible entries by taking different values over K attributes. We define dataset $D(K; m_1, \ldots, m_K)$ as *type I non-symmetric synthetic dataset*. Hence, given the input parameters $(K; m_1, \ldots, m_K)$, we obtain the master table of type I non-symmetric synthetic dataset with the number of rows $\Pi_{k=1}^{K} m_k$.

For example, assume dataset $D(3; 2, 3, 4)$ has 3 attributes A, B and C. Attribute A takes two values 0 or 1; Attribute B takes three values 0, 1 or 2; Attribute C takes four values 0, 1, 2, or 3. The master table has 24 (= $2 \times 3 \times 4$) entries. And the master table is shown in Table 6.

Note that we could obtain the input parameters $(K; m_1, \ldots, m_K)$ for a type I non-symmetric dataset based on a symmetric synthetic dataset $D(K, m)$ by randomly choosing a number m_k from $\{1, 2, \ldots, m\}$ as the number of values for each attribute a_k.

3.1.2 Views

The non-symmetric dataset $D(K; m_1, \ldots, m_K)$ has 2^K different associated views. We examine a relationship between the size of view and the number of values in each of its attribute. The size of view $\{a_{k_1}, a_{k_2}, \ldots, a_{k_l}\}$, measured by its number of rows, is $m_{k_1} m_{k_2} \cdots m_{k_l}$.

The views in a type I non-symmetric dataset lacks of symmetric properties while it has its own properties. Let $m_{\min} = \min_{1 \le k \le K} m_k$. Hence, if V_1 is a descendent of V_2 in the view lattice of $D(K; m_1, \ldots, m_K)$, V_2 is at least $m_{\min}$ times the size of V_1. And identical number of attributes in different views does not guarantee identical size of them.

For example, $D(3; 2, 3, 4)$ has 8 ($= 2^3$) views. As shown in Table 7, view $\{A\}$, $\{B\}$ and $\{C\}$ has 2, 3 and 4 rows, respectively. And view $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$ has 6 ($= 2 \times 3$), 8 ($= 2 \times 4$) and 12 ($= 3 \times 4$) rows, respectively, shown in Table 8. The raw-data view has 24 rows. View $\{A, B\}$ is three times the size of view $\{A\}$.

The size of views in $D(4; 5, 2, 7, 11)$ and $D(5; 6, 8, 5, 13, 7)$ are shown in Table 9 and Table 10, respectively.

3.2 Type II non-symmetric synthetic dataset

3.2.1 Construction

The type I non-symmetric dataset lacks of symmetric properties while it has its own special structure in the master table and views. We consider to break such properties in a new dataset, so-called *type II non-symmetric synthetic dataset*, by partially eliminating rows from the master table of a type I non-symmetric synthetic dataset obtained in the previous section.

The easy way to do the elimination is to randomly eliminating each row in a type I non-symmetric synthetic dataset with a certain probability. However, after conducting this elimination on some datasets, we observe that the size of views in the new datasets does not change much. Thus, we derive the following elimination procedure for obtaining a type II non-symmetric synthetic dataset.

Given a type I non-symmetric synthetic dataset $D(K; m_1, \ldots, m_K)$ with attributes $a_1, \ldots, a_K$, we conduct an elimination process based from attribute a_1 to a_K. To do this, for each attribute a_k, the sub-elimination process consists of two steps. We first do elimination from the first row of the master table to the end. Assume each row in the master table of $D(K; m_1, \ldots, m_K)$ would be kept with an identical probability p_k, and equivalently, eliminated with probability $q_k = 1 - p_k$. Secondly, for each eliminated row r, we also eliminate the rows in the master table with the same values as r on all attributes except a_k.

Let us denote by S_0 the number of rows in the master table of $D(K; m_1, \ldots, m_K)$. $S_0 = \prod_{k=1}^{K} m_k$. In order to evaluate the expected number of rows remaining in the master table after elimination, we first reorder all the entries in the master table grouped by the attributes $a_2, \ldots, a_K$. Equivalently, we divide the entries in the master table into $\prod_{k=2}^{K} m_k$ groups, such that the rows in each group have identical values over $a_2, \ldots, a_K$, and are different from each other only on the value of attribute a_1. In other words, each group corresponds to one row in view $\{a_2, a_3, \ldots, a_K\}$. And the number of rows in each group is m_1, which refers to the number of values on a_1. If one row is eliminated at the first step of the sub-elimination process, all the rows in its group are eliminated at the second step. In other words, each group of rows will be eliminated or remaining in the master table simultaneously. And the probability of eliminating the whole group of rows is $1 - p_1^{m_1}$. Thus, the expected number of rows in each group remaining in the master table is $m_1 p_1^{m_1}$. And the expected number of rows remaining in the master table after the elimination based on attribute a_1, denoted by S_1, is

$$S_1 = \sum_{group} m_1 p_1^{m_1} = m_2 \cdots m_K m_1 p_1^{m_1} = S_0 p_1^{m_1}$$

To evaluate the expected remaining rows after elimination based on attribute a_2, we reorder the remaining entries after elimination based on a_1 grouped by a_1, a_3, $a_4, \ldots, a_K$. The S_1 rows are divided into $m_1 m_3 \cdots m_K$ groups. The number of rows in group g, denoted by m_g, is no more than m_2, i.e., $m_g \leq m_2, \forall g$. Thus, the probability of eliminating all the rows in group g is $1 - p_2^{m_g}$. And the expected number of rows remaining in the master table after the elimination based on attribute a_2, denoted by S_2, is

$$S_2 = \sum_{g=1}^{m_1 m_3 \cdots m_K} m_g p_2^{m_g} \geq \sum_{g=1}^{m_1 m_3 \cdots m_K} m_g p_2^{m_2} = S_1 p_2^{m_2}$$

Thus,

$$S_2 \geq S_0 p_1^{m_1} p_2^{m_2}$$

.

Similarly, we obtain the expected number after the sub-elimination process based on attribute a_k, denoted by S_k, and

$$S_k \geq S_0 p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$$

And thus, the expected number of remaining rows after the whole elimination process, denoted by $\bar{S}$, satisfies

$$\bar{S} \geq S_0 \prod_{k=1}^{K} p_k^{m_k} = \prod_{k=1}^{K} m_k p_k^{m_k} \tag{1}$$

As a result, given the lower bound of the expected rows, after elimination in the master table, denoted by S_L, we could choose the elimination probability q_k for each sub-elimination process based on attribute a_k. The algorithm of obtaining a type II non-symmetric synthetic dataset D' by partial elimination of the rows in a given type

I non-symmetric synthetic dataset $D(K; m_1, \ldots, m_K)$ is shown as follows.

```
Step 0. Input $S_L$ and $D(K; m_1, \ldots, m_K)$. Choose $p_1, \ldots, p_K$ such that
    $S_L \geq \prod_{i=1}^{K} m_i p_i^{m_i}$. Set $k = 1$.
Step 1. Mark each row in the current table as 'selected' with probability $1 - p_k$.
    (If a row is marked, it will be eliminated.)
Step 2. Order the rows in the current table grouped by attribute $a_1, a_2, \ldots, a_{k-1}$,
    $a_{k+1}, \ldots, a_K$.
    For $g = 1$ to $\prod_{i=1}^{K} m_i / m_k$
        In each group $g$, if there exists one row marked, then mark all the $m_k$ rows
        in that group.
Step 3. Eliminate all the marked rows in the table. If $k = K$, output the table as
    $D'$, otherwise $k = k + 1$ and go to step 1.
```

3.2.2 Views

After the new master table is obtained, all the views will be redefined. Thus, we could not guarantee that the views with identical number of attributes have identical size. And the relationship between the size of view and the number of values in each of its attribute is no longer established.

4 Conclusions

In this report, we have defined a symmetric synthetic dataset and two kinds of non-symmetric synthetic dataset. The views of the symmetric synthetic datasets have some symmetric properties while the non-symmetric dataset may not. Once the input parameters for dataset construction are given, the type I non-symmetric synthetic dataset is a deterministic set while the type II non-symmetric dataset is a random one. All these datasets are beneficial for testing the algorithms in computational experiments such as models in Asgharzadeh[1].

References

[1] Z. T. Asgharzadeh, R. Chirkova and Y. Fathi, *Exact and Inexact Methods for Solving the Problems of View Selection*, International Journal of Business Intelligence and Data Mining, 2008

Appendix

A	B	C
0	0	0
0	0	1
0	0	2
0	0	3
0	1	0
0	1	1
0	1	2
0	1	3
0	2	0
0	2	1
0	2	2
0	2	3
0	3	0
0	3	1
0	3	2
0	3	3
1	0	0
1	0	1
1	0	2
1	0	3
1	1	0
1	1	1
1	1	2
1	1	3
1	2	0
1	2	1
1	2	2
1	2	3
1	3	0
1	3	1
1	3	2
1	3	3
2	0	0
2	0	1
2	0	2
2	0	3
2	1	0
2	1	1
2	1	2
2	1	3
2	2	0
2	2	1
2	2	2
2	2	3
2	3	0
2	3	1
2	3	2
2	3	3
3	0	0
3	0	1
3	0	2
3	0	3
3	1	0
3	1	1
3	1	2
3	1	3
3	2	0
3	2	1
3	2	2
3	2	3
3	3	0
3	3	1
3	3	2
3	3	3

Table 1: The symmetric dataset $D(3, 4)$

A	B	C
0	0	0
1	1	1
2	2	2
3	3	3

Table 2: View $\{A\}$, $\{B\}$ and $\{C\}$ in the symmetric dataset $D(3, 4)$

A	B
0	0
0	1
0	2
0	3
1	0
1	1
1	2
1	3
2	0
2	1
2	2
2	3
3	0
3	1
3	2
3	3

A	C
0	0
0	1
0	2
0	3
1	0
1	1
1	2
1	3
2	0
2	1
2	2
2	3
3	0
3	1
3	2
3	3

B	C
0	0
0	1
0	2
0	3
1	0
1	1
1	2
1	3
2	0
2	1
2	2
2	3
3	0
3	1
3	2
3	3

Table 3: View $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$ in the symmetric dataset $D(3, 4)$

view	size	view	size	view	size	view	size
(0,0,0,1)	7	(0,0,1,0)	7	(0,1,0,0)	7	(1,0,0,0)	7
(0,0,1,1)	49	(0,1,0,1)	49	(0,1,1,0)	49	(1,0,0,1)	49
(1,0,1,0)	49	(1,1,0,0)	49				
(0,1,1,1)	343	(1,0,1,1)	343	(1,1,1,0)	343		
(1,1,1,1)	2401						

Table 4: Sizes of views in $D(4, 7)$

view	size	view	size	view	size	view	size	view	size
(0,0,0,0,1)	9	(0,0,0,1,0)	9	(0,0,1,0,0)	9	(0,1,0,0,0)	9	(1,0,0,0,0)	9
(0,0,0,1,1)	81	(0,0,1,0,1)	81	(0,0,1,1,0)	81	(0,1,0,0,1)	81	(0,1,0,1,0)	81
(0,1,1,0,0)	81	(1,0,0,0,1)	81	(1,0,0,1,0)	81	(1,0,1,0,0)	81	(1,1,0,0,0)	81
(0,0,1,1,1)	729	(0,1,0,1,1)	729	(0,1,1,0,1)	729	(0,1,1,1,0)	729	(1,0,0,1,1)	729
(1,0,1,0,1)	729	(1,0,1,1,0)	729	(1,1,0,0,1)	729	(1,1,0,1,0)	729	(1,1,1,0,0)	729
(0,1,1,1,1)	6561	(1,0,1,1,1)	6561	(1,1,0,1,1)	6561	(1,1,1,0,1)	6561	(1,1,1,1,0)	6561
(1,1,1,1,1)	59049								

Table 5: Sizes of views in $D(5, 9)$

A	B	C
0	0	0
0	0	1
0	0	2
0	0	3
0	1	0
0	1	1
0	1	2
0	1	3
0	2	0
0	2	1
0	2	2
0	2	3
1	0	0
1	0	1
1	0	2
1	0	3
1	1	0
1	1	1
1	1	2
1	1	3
1	2	0
1	2	1
1	2	2
1	2	3

Table 6: Non-symmetric dataset $D(3; 2, 3, 4)$

A
0
1

B
0
1
2

C
0
1
2
3

Table 7: View $\{A\}$, $\{B\}$ and $\{C\}$ in the non-symmetric dataset $D(3; 2, 3, 4)$

A	B
0	0
0	1
0	2
1	0
1	1
1	2

A	C
0	0
0	1
0	2
0	3
1	0
1	1
1	2
1	3

B	C
0	0
0	1
0	2
0	3
1	0
1	1
1	2
1	3
2	0
2	1
2	2
2	3

Table 8: View $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$ in the non-symmetric dataset $D(3; 2, 3, 4)$

view	size	view	size	view	size	view	size
(0,0,0,1)	5	(0,0,1,0)	2	(0,1,0,0)	7	(1,0,0,0)	11
(0,0,1,1)	10	(0,1,0,1)	35	(0,1,1,0)	14	(1,0,0,1)	55
(1,0,1,0)	22	(1,1,0,0)	77				
(0,1,1,1)	70	(1,0,1,1)	110	(1,1,1,0)	154		
(1,1,1,1)	770						

Table 9: Sizes of views in $D(4; 5, 2, 7, 11)$

view	size	view	size	view	size	view	size	view	size
(0,0,0,0,1)	6	(0,0,0,1,0)	8	(0,0,1,0,0)	5	(0,1,0,0,0)	13	(1,0,0,0,0)	7
(0,0,0,1,1)	48	(0,0,1,0,1)	30	(0,0,1,1,0)	40	(0,1,0,0,1)	78	(0,1,0,1,0)	104
(0,1,1,0,0)	65	(1,0,0,0,1)	42	(1,0,0,1,0)	56	(1,0,1,0,0)	35	(1,1,0,0,0)	91
(0,0,1,1,1)	240	(0,1,0,1,1)	624	(0,1,1,0,1)	390	(0,1,1,1,0)	520	(1,0,0,1,1)	336
(1,0,1,0,1)	210	(1,0,1,1,0)	280	(1,1,0,0,1)	546	(1,1,0,1,0)	728	(1,1,1,0,0)	455
(0,1,1,1,1)	3120	(1,0,1,1,1)	1680	(1,1,0,1,1)	4368	(1,1,1,0,1)	2730	(1,1,1,1,0)	3640
(1,1,1,1,1)	21840								

Table 10: Sizes of views in $D(5; 6, 8, 5, 13, 7)$