

Clustering Via Connections

by
Jonathan McFadden

6/13/2016 -

Table of Contents

| | | |
|-----------------------------|-------|------|
| <u>1. Definitions</u> | 1- | 6/14 |
| • 1.1 - Input Matrix | 1-2 | 6/14 |
| • 1.2 - SVD Matrices | 2-4 | 6/14 |
| • 1.3 - Connection Matrices | 4-8 | 6/14 |
| • 1.4 - Relation sets | 8-12 | 6/14 |
| • 1.5 - Cluster sets | 12-15 | 6/14 |

1 Definitions

1.1 - Input Matrix

Let $\bar{L} \in \mathbb{R}^{m \times n}$ be a matrix that represents how m terms describe a collection of n objects. We express this formally as either

$$L_{ij} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ term describes the } j^{\text{th}} \text{ object.} \\ 0 & \text{otherwise} \end{cases}$$

or

$$L_{ij} = \begin{cases} n & \text{if the } i^{\text{th}} \text{ term describes the } j^{\text{th}} \text{ object } n \text{ times} \\ 0 & \text{otherwise} \end{cases}$$

In both definitions for the L_{ij} we require that $i \in [1, m] \subset \mathbb{Z}^+$ and that $j \in [1, n] \subset \mathbb{Z}^+$.

1.2 - SVD Matrices

We will be using both the full SVD of \bar{L} and the k order reduced SVD of \bar{L} . The order of the reduced SVD, k , must satisfy

$$k < r$$

where $r = \text{RANK}[\bar{L}]$.

1.2.1 - Full SVD

Let the full singular value decomposition of \bar{E} be given by

$$\bar{E} = \bar{U} \bar{S} \bar{V}^T$$

where $\bar{U} \in \mathbb{R}^{m \times r}$, $\bar{S} \in \mathbb{R}^{r \times r}$, and $\bar{V} \in \mathbb{R}^{n \times r}$, with r defined as above. We call \bar{U} the 'row matrix' of \bar{E} , \bar{S} the 'singular value matrix' of \bar{E} , and \bar{V} the 'column matrix' of \bar{E} .

1.2.2 - Reduced SVD

Let the K order reduced singular

value decomposition of \bar{L} be given by

$$\bar{L} \cong \bar{U}_x \bar{S}_x \bar{V}_x^\top$$

where $\bar{U}_x \in \mathbb{R}^{m \times X}$, $\bar{S}_x \in \mathbb{R}^{X \times X}$, and $\bar{V}_x \in \mathbb{R}^{n \times X}$ with X defined as above. We call \bar{U}_x the 'reduced row matrix' of \bar{L} , \bar{S}_x the 'reduced singular value matrix' of \bar{L} , and \bar{V}_x the 'reduced column matrix' of \bar{L} .

1.3 - Connection Matrices

From \bar{L} , we will create two

'Connection Matrices'. These matrices will be used to provide information for clustering either the rows of \bar{L} or the columns of \bar{L} .

1.3.1 - Row Connection Matrix

The first of the 'Connection Matrices' is the 'Row Connection Matrix'. This matrix will be used to provide information for clustering the rows of \bar{L} .

We will denote this matrix by $\bar{A}_r \in \mathbb{R}^{m \times m}$ and define it, in terms of \bar{L} , by either

$$\bar{\bar{A}}_v = \bar{\bar{L}} \bar{\bar{L}}^T$$

or

$$(\bar{\bar{A}}_v)_{ij} = \sum_{k=1}^n \left\{ S^*(\bar{\bar{L}}, i, k) \bar{\bar{L}}_{jk} \right\}$$

where $i, j \in [1, m] \subset \mathbb{Z}^+$ and the function $S^*(\bar{\bar{L}}, l, \lambda)$ is defined

$$S^*(\bar{\bar{L}}, l, \lambda) = \begin{cases} 1 & \text{if } \bar{\bar{L}}_{l\lambda} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

with $l \in [1, m] \subset \mathbb{Z}^+$ and $\lambda \in [1, n] \subset \mathbb{Z}^+$.
for the definition of $S^*(\bar{\bar{L}}, l, \lambda)$.

1.3.2 - Column Connection Matrix

The second 'Connection Matrix' is the 'Column Connection Matrix'. This matrix will provide information when clustering the columns of \bar{L} and is denoted $\bar{\bar{A}}_v \in \mathbb{R}^{n \times n}$. We define this matrix, in terms of \bar{L} , by either

$$\bar{\bar{A}}_v = \bar{L}^T \bar{L}$$

or

$$(\bar{\bar{A}}_v)_{ij} = \sum_{k=1}^m \left\{ S^*(\bar{L}_{,k}, :)^T \bar{L}_{kj} \right\}$$

where $i, j \in [1, n] \subset \mathbb{Z}^+$ and $S^*(\bar{E}, \lambda, \gamma)$ is as defined above.

1.4 - Relation Sets

We will now define two sets of 2-tuples, with one for rows and the other for columns. These sets will describe the relation or relations any row/column element with any other row/column element as the strength of the relation or relations in question.

1.4.1 - Row Relation Set

Let \mathbb{U} be a set of m 2-tuples such that

$$\mathbb{U} = \left\{ \{M_i, U_i\}, \{M_2, U_2\}, \dots, \{M_m, U_m\} \right\}$$

where the $\{M_i, U_i\}$ are 2-tuples representing all of the rows to which the i^{th} row is related, U_i , and the strength of those relations, M_i ; with $i \in [1, m] \subset \mathbb{Z}^+$. Let us now define U_i and M_i concurrently as

$$U_i = \left\{ j \in [1, m] \subset \mathbb{Z}^+ \mid (\bar{A}_0)_{ij} \neq \emptyset \right\}$$

and

$$\mathbb{M}_i = \left\{ (\bar{A}_0)_{ij} \mid j \in [1, m] \subset \mathbb{Z}^+ \text{ and } (\bar{A}_0)_{ij} \neq 0 \right\}$$

These definitions require that

$$|\mathbb{M}_i| = |\mathbb{U}_i|$$

for all $i \in [1, m] \subset \mathbb{Z}^+$. Additionally, the strength of the connection to the l^{th} element in \mathbb{U}_i is represented by the l^{th} element in \mathbb{M}_i .

1.4.2 - Column Relation Set

Let \mathbb{W} be a set of n 2-tuples such that

$$\mathbb{V} = \{\{N_1, V_1\}, \{N_2, V_2\}, \dots, \{N_n, V_n\}\}$$

where the $\{N_j, V_j\}$ are 2-tuples representing all of the columns to which the j^{th} column is related, V_j , and the strength of those relations, N_j ; with $j \in [1, n] \subset \mathbb{Z}^+$. Let us now define V_j and N_j concurrently as

$$V_j = \{i \in [1, n] \subset \mathbb{Z}^+ \mid (\bar{A}_v)_{ji} \neq 0\}$$

and

$$N_j = \{(\bar{A}_v)_{ji} \mid i \in [1, n] \subset \mathbb{Z}^+ \text{ and } (\bar{A}_v)_{ji} \neq 0\}$$

These definitions require that

$$|\mathbb{N}_j| = |\mathbb{V}_j|$$

for all $j \in [1, n] \subset \mathbb{Z}^+$. Additionally, the strength of the connection to the l^{th} element in \mathbb{V}_j is represented by the l^{th} element in \mathbb{N}_j .

1.5 - Cluster Sets

We now define two sets of sets, with one set representing the row clusters and the other representing the column clusters.

These sets consist of sets that represent the members of each cluster.

1.5.1 - Row Cluster Set

The 'Row Cluster Set' is a set of u_m sets representing each of the u_m clusters. We denote this \mathbb{U}^* and describe it by

$$\mathbb{U}^* = \{ U_1^*, U_2^*, \dots, U_{u_m}^* \}$$

where the U_μ^* are sets containing the members of the μ^{th} cluster, with $\mu \in [1, u_m] \subset \mathbb{Z}^+$. Each U_μ^* is constructed from the $\{M_i, \mathbb{U}_i\}$ of

its (the V_v^*) constituent elements,
via a process which is
defined later.

1.5.2 - Column Cluster Set

The 'Column Cluster Set' is a set of v_n sets representing each of the v_n clusters. We will denote this V^* and describe it by

$$V^* = \{V_1^*, V_2^*, \dots, V_{v_n}^*\}$$

where the V_v^* are sets containing the members of the v^{th} cluster, with $v \in [1, v_n] \subset \mathbb{Z}^+$. Each of the V_v is constructed from

the $\{N_j, V_j\}$ of its (the V_j^*) constituent elements through a process which will be described later.

2 Disjoint Sets

When \bar{L} consists of a series of disjoint sets, the algorithm proceeds by following the steps

- 1) Compute $\bar{A}_o =$
- 2) Construct $U = \{\{M_1, U_1\}, \dots, \{M_m, U_m\}\}$
- 3) Complete the U_i :
- 4) Construct the U^* set
- 5) Compute \bar{A}_v
- 6) Construct $V = \{\{N_1, V_1\}, \dots, \{N_n, V_n\}\}$
- 7) Complete the V_j
- 8) Construct the V^* set
- 9) Reorder \bar{L} into a new matrix, \bar{L}^*

2.1 - Compute $\bar{\bar{A}}_v$

Using the expression

$$\bar{\bar{A}}_v = \bar{\bar{L}} \bar{\bar{L}}^+$$

compute $\bar{\bar{A}}_v \in \mathbb{R}^{m \times m}$. Alternatively,
the expression

$$(\bar{\bar{A}}_v)_{ij} = \sum_{k=1}^n \left\{ S^*(\bar{\bar{L}}, i, k) (\bar{\bar{L}})_{jk} \right\}$$

may be used if different
weighting is desired.

2.2 - Construct \mathcal{V}

We initially construct \mathcal{V} by

the following routine

```

FOR ( $i = 1:m$ ) {
     $U_i = \emptyset$ ;  $M_i = \emptyset$ ;
    FOR ( $j = 1:m$ ) {
        IF ( $(\bar{A}_v)_{ij} \neq 0$ ) {
             $U_i = U_i \cap \{j\}$ ;
             $M_i = M_i \cap \{(\bar{A}_v)_{ij}\}$ ;
        }
    }
     $U[i, 1] = M_i$ ;  $U[i, 2] = U_i$ ;
}

```

After completion of this routine, we make a copy of this initial U , denoted U_0 .

2.3 - Complete the U Sets

The elements of \mathbb{U} must now be 'completed' to include indirectly related elements. Taking advantage of the disjointedness of \bar{L} , this 'completion' is accomplished using the simple routine below

BOOLEAN ISCHANGED = TRUE;

WHILE (ISCHANGED) {
 ISCHANGED = FALSE;

FOR ($i = 1 : m$) {
 $U_i = U[i, 2];$
 $M_i = U[i, 1];$

FOR ($j = 1 : m$) {
 $U_j = U[j, 2];$
 $M_j = U[j, 1];$

```

|   |
|   | IF ( $U_i \neq U_j$  AND  $U_i \cap U_j \neq \emptyset$ ) {
|   |   |
|   |   |  $U_i = U_i \cup \{U_j \setminus U_i\};$ 
|   |   |
|   |   |  $M_i = M_i \cup \{M_j \setminus M_i\};$ 
|   |   |
|   |   |  $U_j = U_j \cup \{U_i \setminus U_j\};$ 
|   |   |
|   |   |  $M_j = M_j \cup \{M_i \setminus M_j\};$ 
|   |   |
|   |   | ISCHANGED = TRUE;
|   |   |
|   |   | } END
|   |
|   | IF (ISCHANGED) {
|   |   |
|   |   |  $U[j, 2] = U_j;$ 
|   |   |  $U[j, 1] = M_j;$ 
|   |   |
|   |   | } END
|   | } END
|   |
|   | IF (ISCHANGED) {
|   |   |
|   |   |  $U[i, 2] = U_i;$ 
|   |   |  $U[i, 1] = M_i;$ 
|   |   |
|   |   | } END
|   | } END
|   |
|   | } END

```

2.4 - Construct the U^* sets

To construct the U^* sets we look for unique U_i in U and store each unique U_i as its own set in U^* . This is done via the following routine

```
U*[1] = U[1,2]; CLUSTCNT = 1;  
FOR (i=2:m) {  
    Ui = U[i,2]; BOOLEAN ISNEW = TRUE;  
    FOR (j=1:CLUSTCNT) {  
        U*j = U*[j];  
        IF (U*j == Ui) {  
            ISNEW = FALSE;  
            BREAK;  
        } END  
    } END  
  
    IF (ISNEW) {  
        CLUSTCNT = CLUSTCNT + 1;  
        U*[CLUSTCNT] = Ui;  
    } END  
} END
```

2.5 - Compute \bar{A}_v

Using the expression

$$\bar{\bar{A}}_v = \bar{\bar{L}}^T \bar{\bar{L}}$$

compute $\bar{\bar{A}}_v \in \mathbb{R}^{n \times n}$. Alternatively,
the expression

$$(\bar{\bar{A}}_v)_{ij} = \sum_{k=1}^m \left\{ S^*(\bar{\bar{L}}, k, i) (\bar{\bar{L}})_{kj} \right\}$$

may be used if different
weighting is desired.

2.6 - Construct V

We initially construct V by the