Clusters

Sumit Mishra

Department of Computer Science & Engineering
Indian Institute of Information Technology Guwahati

Place: IIT Patna
Date: 4 April 2019

Outline

Outline

1. Clusters
   - Number of Clustering Results
   - How to Obtain
   - Integer Partition and Clustering Results
2. Multi-Objective Based Clustering in Entity Matching
   - Motivation & Problem Statement
   - Basics
   - Work Flow

Number of Clustering Results Obtained from 2 Records

Given 2 points/records, number of possible clustering results?

Number of Clustering Results Obtained from 2 Records

Given 2 points/records, number of possible clustering results?

For 2 records

1. $\{\{r_1, r_2\}\}$
2. $\{\{r_1\}, \{r_2\}\}$

Number of Clustering Results Obtained from 3 Records

Given 3 points/records, number of possible clustering results?

Number of Clustering Results Obtained from 3 Records

Given 3 points/records, number of possible clustering results?

For 3 records

1. $\{\{r_1, r_2, r_3\}\}$
2. $\{\{r_1\}, \{r_2, r_3\}\}$
3. $\{\{r_2\}, \{r_1, r_3\}\}$
4. $\{\{r_3\}, \{r_1, r_2\}\}$
5. $\{\{r_1\}, \{r_2\}, \{r_3\}\}$

Number of Clustering Results Obtained from 4 Records

Given 4 points/records, number of possible clustering results?

Number of Clustering Results Obtained from 4 Records

Given 4 points/records, number of possible clustering results?

For 4 records

1. $\{\{r_1, r_2, r_3, r_4\}\}$
2. $\{\{r_1, r_2\}, \{r_3, r_4\}\}$
3. $\{\{r_1, r_3\}, \{r_2, r_4\}\}$
4. $\{\{r_1, r_4\}, \{r_2, r_3\}\}$
5. $\{\{r_1\}, \{r_2, r_3, r_4\}\}$
6. $\{\{r_2\}, \{r_1, r_3, r_4\}\}$
7. $\{\{r_3\}, \{r_1, r_2, r_4\}\}$
8. $\{\{r_4\}, \{r_1, r_2, r_3\}\}$
9. $\{\{r_1\}, \{r_2\}, \{r_3, r_4\}\}$
10. $\{\{r_1\}, \{r_3\}, \{r_2, r_4\}\}$
11. $\{\{r_1\}, \{r_4\}, \{r_2, r_3\}\}$
12. $\{\{r_2\}, \{r_3\}, \{r_1, r_4\}\}$
13. $\{\{r_2\}, \{r_4\}, \{r_1, r_3\}\}$
14. $\{\{r_3\}, \{r_4\}, \{r_1, r_2\}\}$
15. $\{\{r_1\}, \{r_2\}, \{r_3\}, \{r_4\}\}$

Number of Clustering Results Obtained from n Records

Given n points/records, number of possible clustering results?

Number of Clustering Results Obtained from n Records

Given n points/records, number of possible clustering results?

Using Bell number [1]

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k, \qquad B_0 = B_1 = 1 \tag{1}$$

Number of Clustering Results Obtained from n Records

Given n points/records, number of possible clustering results?

Using Bell number [1]

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k, \qquad B_0 = B_1 = 1 \tag{1}$$

Using Stirling numbers of the second kind [8]

$$B_n = \sum_{k=0}^{n} \left\{ {n \atop k} \right\} \tag{2}$$

Here, the Stirling number $\left\{ {n \atop k} \right\}$ is the number of ways to partition a set of cardinality n into exactly k nonempty subsets.

Bell Triangle

1									
1	**2**								
2	3	**5**							
5	7	10	**15**						
15	20	27	37	**52**					
52	67	87	114	151	**203**				
203	255	322	409	523	674	**877**			
877	1080	1335	1657	2066	2589	3263	**4140**		
4140	5017	6097	7432	9089	11155	13744	17007	**21147**	
21147	25287	30304	36401	43833	52922	64077	77821	94828	**115975**

Figure 1 : Bell triangle for 10 records.

Outline

Arrangements of n Records in Clustering Results

Given n points/records, how the records are arranged in the clustering result?

Arrangements of n Records in Clustering Results

Given n points/records, how the records are arranged in the clustering result? Bell Polynomial [1, 2]

$$B_n(x_1, \dots, x_n) = \sum_{k=1}^{n} B_{n,k}(x_1, x_2, \dots, x_{n-k+1}) \tag{3}$$

$B_{n,k}(x_1, x_2, \dots, x_{n-k+1})$ is the partial Bell polynomial and is given by Equation (4).

$$B_{n,k}(x_1, x_2, \dots, x_{n-k+1}) = \sum \frac{n!}{j_1! j_2! \cdots j_{n-k+1}!} \left(\frac{x_1}{1!}\right)^{j_1} \left(\frac{x_2}{2!}\right)^{j_2} \cdots \left(\frac{x_{n-k}}{(n-k)!}\right)^{j_{n-k}} \left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}} \tag{4}$$

where the sum is taken over all sequences $j_1, j_2, j_3, \dots, j_{n-k+1}$ of non-negative integers such that $j_1 + j_2 + \cdots = k$ and $j_1 + 2j_2 + 3j_3 + \cdots = n$.

Arrangements of 3 Records in Clustering Results

3^{rd} complete Bell polynomial is given by Equation (5).

$$\begin{aligned} B_3(x_1, x_2, x_3) &= \sum\nolimits_{k=1}^{3} B_{3,k}(x_1, x_2, \ldots, x_{3-k+1}) \\ &= B_{3,1}(x_1, x_2, x_3) + B_{3,2}(x_1, x_2) + B_{3,3}(x_1) \\ &= (x_3) + (3x_1x_2) + \left(x_1^3\right) \end{aligned} \tag{5}$$

In simple terms Equation (5) can be written as Equation (6).

$$B_3(x) = x + 3x^2 + x^3 \tag{6}$$

Arrangements of 3 Records in Clustering Results . . .

- One way to group 3 records in single cluster,
  1. $\{\{r_1, r_2, r_3\}\}$
- Three ways to group 3 records in 2 clusters,
  1. $\{\{r_1\}, \{r_2, r_3\}\}$
  2. $\{\{r_2\}, \{r_1, r_3\}\}$
  3. $\{\{r_3\}, \{r_1, r_2\}\}$
- One way to group 3 records in 3 clusters.
  1. $\{\{r_1\}, \{r_2\}, \{r_3\}\}$

Arrangements of 4 Records in Clustering Results

4^{th} complete Bell polynomial is given by Equation (7).

$$\begin{aligned} B_4(x_1, x_2, x_3, x_4) &= \sum_{k=1}^{4} B_{4,k}(x_1, x_2, \ldots, x_{4-k+1}) \\ &= B_{4,1}(x_1, x_2, x_3, x_4) + B_{4,2}(x_1, x_2, x_3) + B_{4,3}(x_1, x_2) + B_{4,4}(x_1) \\ &= (x_4) + (3x_2^2 + 4x_1x_3) + (6x_1^2x_2) + (x_1^4) \end{aligned} \tag{7}$$

In simple terms Equation (7) can be written as Equation (8).

$$B_4(x) = x + 7x^2 + 6x^3 + x^4 \tag{8}$$

Arrangements of 4 Records in Clustering Results . . .

- One way to group 4 records in single cluster,
  1. $\{\{r_1, r_2, r_3, r_4\}\}$
- Seven $(3+4)$ ways to group 4 records in 2 clusters,
  1. $\{\{r_1, r_2\}, \{r_3, r_4\}\}$
  2. $\{\{r_1, r_3\}, \{r_2, r_4\}\}$
  3. $\{\{r_1, r_4\}, \{r_2, r_3\}\}$
  4. $\{\{r_1\}, \{r_2, r_3, r_4\}\}$
  5. $\{\{r_2\}, \{r_1, r_3, r_4\}\}$
  6. $\{\{r_3\}, \{r_1, r_2, r_4\}\}$
  7. $\{\{r_4\}, \{r_1, r_2, r_3\}\}$
- Six ways to group 4 records in 3 clusters,
  1. $\{\{r_1\}, \{r_2\}, \{r_3, r_4\}\}$
  2. $\{\{r_1\}, \{r_3\}, \{r_2, r_4\}\}$
  3. $\{\{r_1\}, \{r_4\}, \{r_2, r_3\}\}$
  4. $\{\{r_2\}, \{r_3\}, \{r_1, r_4\}\}$
  5. $\{\{r_2\}, \{r_4\}, \{r_1, r_3\}\}$
  6. $\{\{r_3\}, \{r_4\}, \{r_1, r_2\}\}$
- One way to group 4 records in 4 clusters.
  1. $\{\{r_1\}, \{r_2\}, \{r_3\}, \{r_4\}\}$

Outline

1. Clusters
   - Number of Clustering Results
   - How to Obtain
   - Integer Partition and Clustering Results
2. Multi-Objective Based Clustering in Entity Matching
   - Motivation & Problem Statement
   - Basics
   - Work Flow

Integer Partition

A partition of a positive integer n is defined to be a sequence of positive integers whose sum is n [3].

Integer partition for $n = 3$

1. 3
2. $1 + 2$
3. $1 + 1 + 1$

Integer partition for $n = 4$

1. 4
2. $2 + 2$
3. $1 + 3$
4. $1 + 1 + 2$
5. $1 + 1 + 1 + 1$

Integer Partition . . .

An asymptotic expression for number of partitions of an integer n is given by Equation (9) [5, 7].

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right) \text{ as } n \to \infty. \tag{9}$$

Integer Partition and Clustering Results for 3 Records

Serial No.	Number of Cluster	Possible Partition	Partial Bell Polynomial	No. of Clustering Results Corresponding to a Partition
1	1	$\{3\}$	x_3	1
2	2	$\{1,2\}$	$3x_1x_2$	3
3	3	$\{1,1,1\}$	x_1^3	1
Total number of clustering results B_3				**5**

Table 1 : Possible partition of 3 records along with the number of clustering results corresponding to each partition.

Integer Partition and Clustering Results for 4 Records

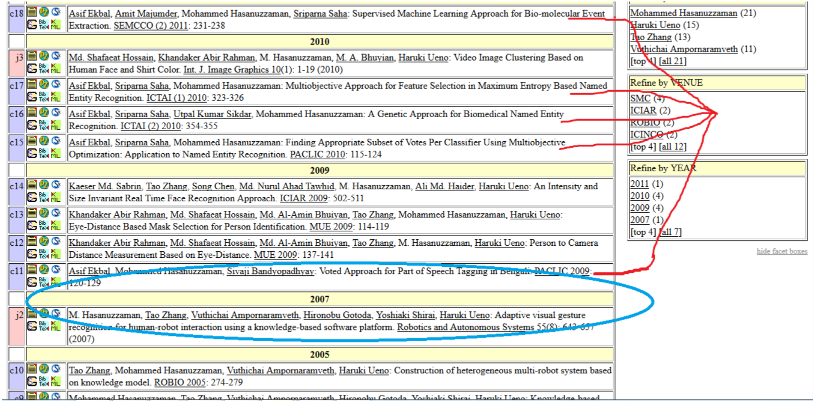
Serial No.	Number of Cluster	Possible Partition	Partial Bell Polynomial	No. of Clustering Results Corresponding to a Partition
1	1	$\{4\}$	x_4	1
2	2	$\{2,2\}$	$3x_2^2 + 4x_1x_2$	3
3		$\{1,3\}$		4
4	3	$\{1,1,2\}$	$6x_1^2x_2$	6
5	4	$\{1,1,1,1\}$	x_1^4	1
Total number of clustering results B_4				**15**

Table 2 : Possible partition of 4 records along with the number of clustering results corresponding to each partition.

Outline

1. Clusters
   - Number of Clustering Results
   - How to Obtain
   - Integer Partition and Clustering Results
2. Multi-Objective Based Clustering in Entity Matching
   - Motivation & Problem Statement
   - Basics
   - Work Flow

Motivation



c18	Asif Ekbal, Amit Majumder, Mohammed Hasanuzzaman, Sriparna Saha: Supervised Machine Learning Approach for Bio-molecular Event Extraction. SEMCCO (2) 2011: 231-238
	2010
j3	Md. Shafaeat Hossain, Khandaker Abir Rahman, M. Hasanuzzaman, M. A. Bhuyian, Haruki Ueno: Video Image Clustering Based on Human Face and Shirt Color. Int. J. Image Graphics 10(1): 1-19 (2010)
c17	Asif Ekbal, Sriparna Saha, Mohammed Hasanuzzaman: Multiobjective Approach for Feature Selection in Maximum Entropy Based Named Entity Recognition. ICTAI (1) 2010: 323-326
c16	Asif Ekbal, Sriparna Saha, Utpal Kumar Sikdar, Mohammed Hasanuzzaman: A Genetic Approach for Biomedical Named Entity Recognition. ICTAI (2) 2010: 354-355
c15	Asif Ekbal, Sriparna Saha, Mohammed Hasanuzzaman: Finding Appropriate Subset of Votes Per Classifier Using Multiobjective Optimization: Application to Named Entity Recognition. PACLIC 2010: 115-124
	2009
c14	Kaeser Md. Sabrin, Tao Zhang, Song Chen, Md. Nurul Ahad Tawhid, M. Hasanuzzaman, Ali Md. Haider, Haruki Ueno: An Intensity and Size Invariant Real Time Face Recognition Approach. ICIAR 2009: 502-511
c13	Khandaker Abir Rahman, Md. Shafaeat Hossain, Md. Al-Amin Bhuiyan, Tao Zhang, Mohammed Hasanuzzaman, Haruki Ueno: Eye-Distance Based Mask Selection for Person Identification. MUE 2009: 114-119
c12	Khandaker Abir Rahman, Md. Shafaeat Hossain, Md. Al-Amin Bhuiyan, Tao Zhang, M. Hasanuzzaman, Haruki Ueno: Person to Camera Distance Measurement Based on Eye-Distance. MUE 2009: 137-141
c11	Asif Ekbal, Mohammed Hasanuzzaman, Sivaji Bandyopadhyay: Voted Approach for Part of Speech Tagging in Bengali. PACLIC 2009: 120-129
	2007
j2	M. Hasanuzzaman, Tao Zhang, Vuthichai Ampornaramveth, Hironobu Gotoda, Yoshiaki Shirai, Haruki Ueno: Adaptive visual gesture recognition for human-robot interaction using a knowledge-based software platform. Robotics and Autonomous Systems 55(8): 643-657 (2007)
	2005
c10	Tao Zhang, Mohammed Hasanuzzaman, Vuthichai Ampornaramveth, Haruki Ueno: Construction of heterogeneous multi-robot system based on knowledge model. ROBIO 2005: 274-279

Mohammed Hasanuzzaman (21)
Haruki Ueno (15)
Tao Zhang (13)
Vuthichai Ampornaramveth (11)
[top 4] [all 21]

Refine by VENUE
SMC (4)
ICIAR (2)
ROBIO (2)
ICINCO (2)
[top 4] [all 12]

Refine by YEAR
2011 (1)
2010 (4)
2009 (4)
2007 (1)
[top 4] [all 7]

hide facet boxes

Observation

- There are two different authors with the same name.

Problem Statement

Motivation

- DBLP contains more than 1300 papers published by authors having the name "Wei Wang".
- How many authors having the same name?
- What is the categorization of the papers?

DEFINITION: Entity Matching [6]

Given a name 'pName' and a set of records $\mathbb{R} = \{r_1, r_2, \ \ldots, r_n\}$ corresponding to name 'pName', the Entity Matching is to divide the records in $\mathbb{R}$ into different clusters such that the following holds.

- All the records in a cluster belong to an entity.
- All the records by an entity should be in a single cluster.

Outline

1. Clusters
   - Number of Clustering Results
   - How to Obtain
   - Integer Partition and Clustering Results
2. Multi-Objective Based Clustering in Entity Matching
   - Motivation & Problem Statement
   - Basics
   - Work Flow

Evolutionary Algorithm based Approach

Motivation

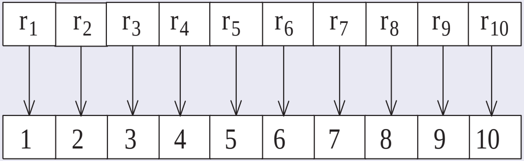
- Modeled entity matching problem as an optimization problem.
- Used evolutionary algorithm as an optimization framework.
- Single as well as multiple objectives are considered.

Algorithm 1 GENETIC ALGORITHM

1: Initialize population P
2: Evaluate the fitness of all individuals
3: Select fitter individuals for reproduction
4: Apply recombination among individuals
5: Mutate individuals
6: Evaluate the fitness of the modified individuals
7: Generate a new population

Chromosome Initialization

Record Encoding



Chromosome Representation

- Represents the representative of the cluster, *i.e.*, one of the elements from the cluster.
- The K length of *Chromosome* means that there are K clusters.
- The elements in the *Chromosome* are distinct.

2	**5**	**8**	**10**

Cluster representative

Population Initialization and Assignment of Records

Population Initialization

- The number of clusters is unknown.
- Size of *chromosome* varies between 2 and $n-1$.
- To initialize population, each *chromosome* in the population is initialized.

Assignment of Records

- $\mathbb{R} = \{r_1, r_2, \ldots, r_{10}\}$
- Chromosome $= \{r_2, r_5, r_8, r_{10}\}$
- $\mathbb{UR} = \{r_1, r_3, r_4, r_6, r_7, r_9\}$
- Distance Measure

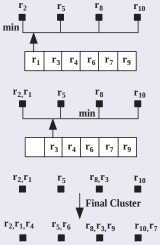


Figure 2 : Assignment of records.

Crossover and Mutation

Crossover Operation on Chromosomes

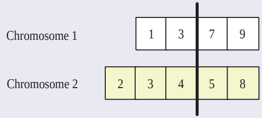


Figure 3 : Before crossover

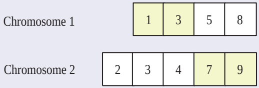


Figure 4 : After crossover

Mutation Operation on Chromosome

Three mutation operations are considered.

i. Insert
ii. Delete
iii. Update

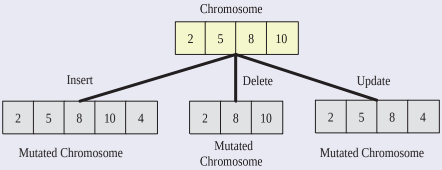


Figure 5 : Mutation operation

Outline

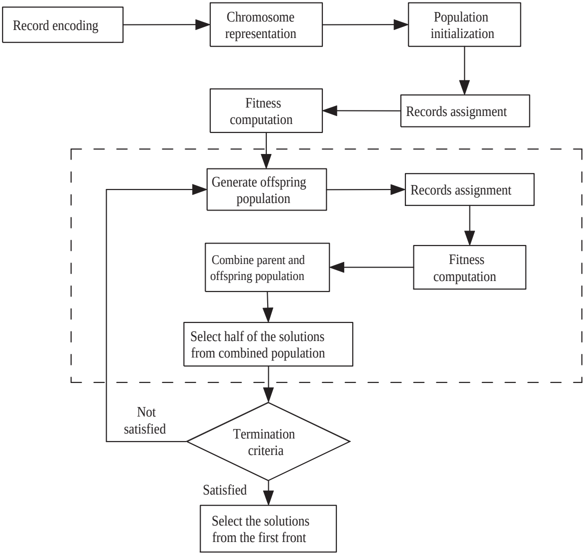


Figure 6 : Flow of MOO-EMT

NSGA-II

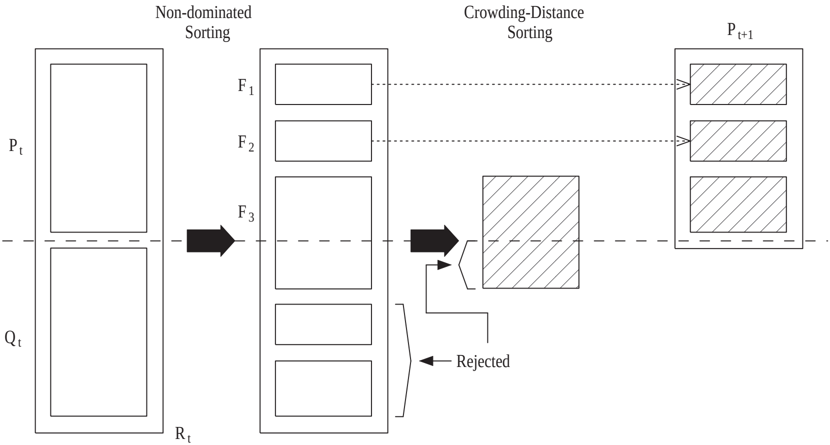


Figure 7 : NSGA-II Procedure [4]

NSGA-II ...

Algorithm 2 NSGA-II

Input: $\mathbb{P}_t$: Population for t^{th} generation
Output: $\mathbb{P}_{t+1}$: Population for $t+1^{th}$ generation

```
Q_t ← Generate offspring population after crossover and mutation operations
R_t ← P_t ∪ Q_t                       // Combine parent and offspring populations
F ← NON-DOMINATED-SORT(R_t)           // F = {F_1, F_2, ..., F_K}, set of
    non-dominated fronts in the decreasing order of their dominance nature
P_{t+1} ← Φ                           // Initialize population for the next generation
k ← 1
while |P_{t+1}| + |F_k| ≤ N do
    P_{t+1} ← P_{t+1} ∪ F_k           // Include k^th non-dominated front in P_{t+1}
    k ← k + 1                         // Check the next front for inclusion in P_{t+1}
// Number of solutions to be included in population P_{t+1}
T = N − (|F_1| + |F_2| + ... + |F_{k−1}|)
CROWDING-DISTANCE-ASSIGNMENT(F_k)     // Calculate crowding distance
    in F_k
Sort the solutions in F_k based on crowding distance
P_{t+1} ← P_{t+1} ∪ F_k [1 : T]       // Choose the first T solutions from F_k
```

NSGA-II . . .

Algorithm 3 $\textsc{Crowding-Distance-Assignment}(I)$

Input: I : Non-dominated front
Output: I : Crowded distance assignment to each solution in I

```
l ← |I|                                   // Number of solutions in non-dominated front I
for i ← 1 to l do
    I[i]_distance ← 0                     // Initialize the crowding distance
for each objective m do
    I ← SORT(I, m)   // Sort the solutions in I in descending order using m^th objective
    I[1]_distance ← ∞                     // Set the value for boundary points
    I[l]_distance ← ∞                     // Set the value for boundary points
    for i ← 2 to l − 1 do
        I[i]_distance ← I[i]_distance + (I[i+1].m − I[i−1].m) / (f_m^max − f_m^min)
```

$$I[i]_{\text{distance}} \leftarrow I[i]_{\text{distance}} + \left(I[i+1].m - I[i-1].m\right) / \left(f_m^{\max} - f_m^{\min}\right)$$

Conclusions

- Discussed the number of clusters and how it can be obtained.
- Also discussed entity matching problem and how it can be solved using multi-objective clustering.

References I

[1] HW Becker and John Riordan.
The arithmetic of bell and stirling numbers.
American journal of Mathematics, pages 385–394, 1948.

[2] Eric Temple Bell.
Partition polynomials.
Annals of Mathematics, pages 38–46, 1927.

[3] David M Burton.
Elementary number theory.
Tata McGraw-Hill Education, 2006.

[4] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan.
A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II.
IEEE Transactions on Evolutionary Computation, 6(2):182–197, April 2002.

[5] Godfrey H Hardy and Srinivasa Ramanujan.
Asymptotic formulaæ in combinatory analysis.
Proceedings of the London Mathematical Society, 2(1):75–115, 1918.

[6] Shaohua Li, Gao Cong, and Chunyan Miao.
Author name disambiguation using a new categorical distribution similarity.
In *Machine learning and knowledge discovery in databases*, pages 569–584. Springer, 2012.

References II

[7] Hans Rademacher.
On the expansion of the partition function in a series.
Annals of Mathematics, pages 416–422, 1943.

[8] Basil Cameron Rennie and Annette Jane Dobson.
On Stirling Numbers of the Second Kind.
Journal of Combinatorial Theory, 7(2):116–121, 1969.

Thank you!