Simultaneous Clustering and Gene Ranking: A Multi-objective Genetic Approach

Kartick Chandra Mondal

Co-authors: Anirban Mukhopadhyay\textsuperscript{2}, Ujjwal Maulik\textsuperscript{3}, Sanghamitra Bandyopadhyay\textsuperscript{4}, Nicolas Pasquier\textsuperscript{1}

\textsuperscript{1}I3S Laboratory (CNRS UMR-6070), University of Nice Sophia-Antipolis, Nice-06108, France, keto004@gmail.com.
\textsuperscript{2}Department of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg-69120, Germany.
\textsuperscript{3}Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
\textsuperscript{4}Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700108, India

CIBB, 16th-18th September 2010
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline

1. Introduction
2. Multi-objective Optimization (MOO)
   - Definition and need of MOO
   - NSGA-II
3. Proposed Algorithm & its Components
4. Experiments and Results
   - Experimental Designs
   - Results and Discussion
5. Conclusions & Feature Scopes
Introduction

Problem Description

Find the classes present in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
Introduction

Problem Description

Find the classes present in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
Introduction

Problem Description

Find the classes presents in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
Introduction

Problem Description

Find the classes present in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
Introduction

Problem Description

Find the classes presents in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
Introduction

Problem Description

Find the classes present in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
Introduction

Problem Description

Find the classes present in the data set and rank the features according to their ability to distinguish those classes, Simultaneously.
What we proposed

1 A multi-objective approach
   • A chromosome representation for the approach.
   • A simultaneous operation for ranking and clustering.

2 R Index for comparing rank.
What we proposed

1. A multi-objective approach
   - A chromosome representation for the approach.
   - A simultaneous operation for ranking and clustering.

2. R Index for comparing rank.
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline

1. Introduction

2. Multi-objective Optimization (MOO)
   - Definition and need of MOO
   - NSGA-II

3. Proposed Algorithm & its Components
   - Proposed Algorithm & its Components

4. Experiments and Results
   - Experimental Designs
   - Results and Discussion

5. Conclusions & Feature Scopes
   - Conclusions & Feature Scopes
Need of MOO

In many real world problems, we have to simultaneously optimize two or more different objectives which are often competitive in nature.

Finding a single solution in these cases is very difficult.

Optimizing each criterion separately may lead to good value of one objective while some unacceptably low value of the other objective(s).
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

Need of MOO

In many real world problems, we have to simultaneously optimize two or more different objectives which are often competitive in nature.

Finding a single solution in these cases is very difficult.

Optimizing each criterion separately may lead to good value of one objective while some unacceptably low value of the other objective(s).
In many real world problems, we have to simultaneously optimize two or more different objectives which are often competitive in nature.

Finding a single solution in these cases is very difficult.

Optimizing each criterion separately may lead to good value of one objective while some unacceptably low value of the other objective(s).
Problem Statement

- Find the vector of the decision variables:
  \[ \overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T \]

- That satisfy \( m \) inequality constraints and \( p \) equality constraints
  \[ g_i(\overline{x}) \geq 0, \quad i = 1, 2, \ldots, m, \]
  \[ h_j(\overline{x}) = 0, \quad j = 1, 2, \ldots, p, \]

- Also optimizes the vector function (consisting of \( k \) objective functions):
  \[ \overline{f}(\overline{x}) = [f_1(\overline{x}), f_2(\overline{x}), \ldots, f_k(\overline{x})]^T. \]
Problem Statement

• Find the vector of the decision variables:

\[ \overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T \]

• That satisfy \( m \) inequality constraints and \( p \) equality constraints

\[ g_i(\overline{x}) \geq 0, \quad i = 1, 2, \ldots, m, \]
\[ h_j(\overline{x}) = 0, \quad j = 1, 2, \ldots, p, \]

• Also optimizes the vector function (consisting of \( k \) objective functions):

\[ \overline{f}(\overline{x}) = [f_1(\overline{x}), f_2(\overline{x}), \ldots, f_k(\overline{x})]^T. \]
Problem Statement

- Find the vector of the decision variables:
  \[ \overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T \]

- That satisfy \( m \) inequality constraints and \( p \) equality constraints
  \[ g_i(\overline{x}) \geq 0, \quad i = 1, 2, \ldots, m, \]
  \[ h_j(\overline{x}) = 0, \quad j = 1, 2, \ldots, p, \]

- Also optimizes the vector function (consisting of \( k \) objective functions):
  \[ \overline{f}(\overline{x}) = [f_1(\overline{x}), f_2(\overline{x}), \ldots, f_k(\overline{x})]^T. \]
Domination Relationship

Let $a$ and $b$ be two solutions. Then $a$ is said to dominate $b$ iff

$$\forall i \in \{1, \ldots, k\}, f_i(b) \leq f_i(a)$$

and

$$\exists j \in \{1, \ldots, k\}, f_j(b) < f_j(a).$$

i.e., for all functions $f_i$, $a$ has a higher or equal value than that of $b$ and also there exists at least one function $f_j$ for which $a$’s value is strictly greater than that of $b$. 

Domination Relation and Pareto-Optimality I
Domination Relation and Pareto-Optimality II

Non-dominated Set

Among a set of solutions $P$, the non-dominated set of solutions $P'$ are those that are not dominated by any solution in the set $P$.

Pareto-optimal Set

The non-dominated set of entire search space $S$ is globally Pareto optimal set.

Pareto Set Approximation

Since it is impossible to identify Pareto-optimal set due to it’s size, identifying the Best-known Pareto set or finding/approximating the Pareto Set is called Pareto Set Approximation.
**Domination Relation and Pareto-Optimality II**

**Non-dominated Set**

Among a set of solutions $P$, the non-dominated set of solutions $P'$ are those that are not dominated by any solution in the set $P$.

**Pareto-optimal Set**

The non-dominated set of entire search space $S$ is globally Pareto optimal set.

**Pareto Set Approximation**

Since it is impossible to identify Pareto-optimal set due to it’s size, identifying the Best-known Pareto set or finding/approximating the Pareto Set is called Pareto Set Approximation.
Domination Relation and Pareto-Optimality II

Non-dominated Set

Among a set of solutions $P$, the non-dominated set of solutions $P'$ are those that are not dominated by any solution in the set $P$.

Pareto-optimal Set

The non-dominated set of entire search space $S$ is globally Pareto optimal set.

Pareto Set Approximation

Since it is impossible to identify Pareto-optimal set due to it’s size, identifying the Best-known Pareto set or finding/approximating the Pareto Set is called Pareto Set Approximation.
Goal of the Optimization

Goal: 1
The best-known Pareto front should be as close as possible to the true Pareto front.

Goal: 2
Solutions in the best-known Pareto set should be uniformly distributed and diverse over of the Pareto front.

Goal: 3
The best-known Pareto front should capture the whole spectrum of the Pareto front.
Goal of the Optimization

Goal: 1
The best-known Pareto front should be as close as possible to the true Pareto front.

Goal: 2
Solutions in the best-known Pareto set should be uniformly distributed and diverse over of the Pareto front.

Goal: 3
The best-known Pareto front should capture the whole spectrum of the Pareto front.
Goal of the Optimization

Goal: 1
The best-known Pareto front should be as close as possible to the true Pareto front.

Goal: 2
Solutions in the best-known Pareto set should be uniformly distributed and diverse over of the Pareto front.

Goal: 3
The best-known Pareto front should capture the whole spectrum of the Pareto front.
Multiobjective Genetic Algorithms

- Non-Pareto approach
  - Vector Evaluated GA (VEGA)
- Pareto-based approach
  - Non-dominated Sorting GA (NSGA and NSGA-II)
  - Niched Pareto GA (NPGA)
  - Strength Pareto Evolutionary Algorithm (SPEA and SPEA2)
  - Pareto Archived Evolutionary Strategy (PAES)
  - Pareto Envelop-based Selection Algorithm (PESA and PESA-II)

Multiobjective Genetic Algorithms

- Non-Pareto approach
  - Vector Evaluated GA (VEGA)
- Pareto-based approach
  - Non-dominated Sorting GA (NSGA and NSGA-II)
  - Niched Pareto GA (NPGA)
  - Strength Pareto Evolutionary Algorithm (SPEA and SPEA2)
  - Pareto Archived Evolutionary Strategy (PAES)
  - Pareto Envelop-based Selection Algorithm (PESA and PESA-II)

Outline

1. Introduction

2. Multi-objective Optimization (MOO)
   - Definition and need of MOO
   - NSGA-II

3. Proposed Algorithm & its Components
   - Proposed Algorithm & its Components

4. Experiments and Results
   - Experimental Designs
   - Results and Discussion

5. Conclusions & Feature Scopes
   - Conclusions & Feature Scopes
NSGA-II Flowchart
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

NSGA-II Flowchart

Non-dominated sorting

P_i

Rank 1
Rank 2
Rank 3
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

NSGA-II Flowchart

\[ P_i \rightarrow \text{Non-dominated sorting} \rightarrow \text{Selection} \rightarrow M_i \]
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

NSGA-II Flowchart

P_i → Non-dominated sorting → Rank 1, Rank 2, Rank 3 → Selection → M_i → Crossover and Mutation → C_i
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

NSGA-II Flowchart
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

NSGA-II Flowchart
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
- Introduction
- Multi-objective Optimization (MOO)
  - Definition and need of MOO
  - NSGA-II
- Proposed Algorithm & its Components
- Experiments and Results
  - Experimental Designs
  - Results and Discussion
- Conclusions & Feature Scopes

NSGA-II Flowchart

Non-dominated sorting

Rank 1
Rank 2
Rank 3

Selection

Crossover and Mutation

Rank 1
Rank 2
Rank 3

Elitism

Rank 4

Non-dominated sorting

P_i

P_i+1

M_i

C_i
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

NSGA-II Flowchart

1. Non-dominated sorting
2. Selection
3. Crossover and Mutation
4. Non-dominated sorting

P_i → Rank 1 → M_i → C_i → P_i+1

Rank 1
Rank 2
Rank 3
Rank 4
Simultaneous Clustering and Gene Ranking
Kartick Chandra Mondal

Outline

1. Introduction
2. Multi-objective Optimization (MOO)
   - Definition and need of MOO
   - NSGA-II
3. Proposed Algorithm & its Components
4. Experiments and Results
   - Experimental Designs
   - Results and Discussion
5. Conclusions & Feature Scopes
Chromosome Representation & Initial Population

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>...</th>
<th>d</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.534</td>
<td>0.8131</td>
<td>.</td>
<td>.</td>
<td>0.1421</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- For Ranking
- For Cluster Center

Single Chromosome

Total Chromosome Length (d + (k x d))
Fitness Computation

- Two validity indices, Xie-Beni and DB, are used here as two objective function.

- Both of these objective functions are of minimization type.
  - Small value of XB index indicates compact and well separated clusters.
  - Small value of DB index indicates the minimum possible similarity to each other clusters.

- After assigning the samples, update the cluster centers and the new cluster centers are replaced in the chromosome.
Fitness Computation

- Two validity indices, Xie-Beni and DB, are used here as two objective function.

- Both of these objective functions are of minimization type.
  - Small value of XB index indicates compact and well separated clusters.
  - Small value of DB index indicates the minimum possible similarity to each other clusters.

- After assigning the samples, update the cluster centers and the new cluster centers are replaced in the chromosome.
Fitness Computation

- Two validity indices, Xie-Beni and DB, are used here as two objective function.

- Both of these objective functions are of minimization type.
  - Small value of XB index indicates compact and well separated clusters.
  - Small value of DB index indicates the minimum possible similarity to each other clusters.

- After assigning the samples, update the cluster centers and the new cluster centers are replaced in the chromosome.
Validity Index: Xie-Beni

The compactness and separation validity function.

The Index

\[ XB = \frac{\pi}{(N \times D_{\text{min}})} \]  (1)

N is the number of points in X.

The Compactness and Separation

\[ \pi = \frac{\sigma_i}{n_i} \]  (2)

\[ D_{\text{min}} = \min \| v_i - v_j \| \]  (3)

\( \sigma_i \) is sum of the squares of fuzzy deviation.
Validity Index: Xie-Beni

The fuzzy deviation

\[ d_{ij} = u_{ij} \| x_j - v_i \| \]  \hspace{1cm} (4)

\( x_j \) is the fuzzy partition of \( X \) where \( (j=1 \text{ to } n) \).
\( v_i \) is \( i_{th} \) cluster center; \( (i=1 \text{ to } n_c) \).
\( u_{ij} \) is the membership value of the data point \( j \) of cluster \( i \).
Validity Index: Davies-Bouldin

The ratio of the sum of within-cluster scatter $S_i$ to between-cluster separation $d_{ij}$.

The Index

$$DB_n = \frac{1}{n} \sum_{i=1}^{n} R_i$$

(5)

Where,

$$R_i = \max_{i,j=1 \text{ to } n, i\neq j} R_{ij}$$

(6)

The Similarity Measure

$R_{ij}$, the similarity measure between the clusters $C_i$ and $C_j$ is defined as

$$R_{ij} = (s_i + s_j)/d_{ij}$$

(7)

Here, $R_{ij}$ is non-negative and symmetric.
Validity Index: Davies-Bouldin

The Scatter
The scatter of cluster $S_i$ is defined as

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} ||x - C_i||$$  \hspace{1cm} (8)

The Distance
The distance or dissimilarity measure between two cluster $C_i$ and $C_j$ is denoted by $d_{ij}$ and is defined as

$$d_{ij} = ||C_i - C_j||$$  \hspace{1cm} (9)
Distance Measure

Weighted Euclidean Distance

Compute the validity indices using weighted distance (weighted euclidean distance).

\[ d(x, y) = \sqrt{\sum_{i=1}^{d} w_i^2 (x_i - y_i)^2} \] (10)
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value \textbf{(Rank)} and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
Non-dominated Sorting

1. It is based on several layers of classifications of the individuals.
2. Non-dominated individuals get a certain dummy fitness value (Rank) and then are removed from the population.
3. Process is repeated until the entire population is classified.
1. If mutation is possible the actual value of the mutate bit is replaced by a random value in the range of [0,1].

2. The same technique is used in both part of the chromosome.
Selection

Use binary tournament selection method with crowded and rank comparison method.
Crowded Binary Tournament Selection

- **Binary Tournament Selection for Single Objective**
  - Select two random solutions (chromosomes) $a$ and $b$ and put the solution $a$ in the mating pool if $a$ has better fitness than $b$.

- **Crowded Binary Tournament Selection**
  - Select two random solutions $a$ & $b$ and put solution $a$ in the mating pool if

    $$\text{Rank}(a) < \text{Rank}(b) \quad \text{or,}$$

    $$\text{Rank}(a) = \text{Rank}(b) \quad \text{and} \quad CD(a) > CD(b)$$
Elitism & Termination

Elitism

1. The generated child population is combined with the parent population of that generation.
2. From this combined population, the non-dominated chromosomes are selected to create a new population for the next generation.

Termination

1. The NSGA-II has been executed for a fixed number of generations.
2. This fixed number is supplied by the user for terminating the process.
Elitism & Termination

Elitism

1. The generated child population is combined with the parent population of that generation.
2. From this combined population, the non-dominated chromosomes are selected to create a new population for the next generation.

Termination

1. The NSGA-II has been executed for a fixed number of generations.
2. This fixed number is supplied by the user for terminating the process.
Final Solution Selection

1. After terminating, the process gives a set of non-dominated solutions in the last generation.

2. The final solution from the non-dominated set is selected through the CP index and R index.
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline

1 Introduction

2 Multi-objective Optimization (MOO)
   Definition and need of MOO
   NSGA-II

3 Proposed Algorithm & its Components
   Proposed Algorithm & its Components

4 Experiments and Results
   Experimental Designs
   Results and Discussion

5 Conclusions & Feature Scopes
   Conclusions & Feature Scopes
Data Preparation

Artificial Data Sets
- Arda25_30_3
- Arda50_75_5

Real Life Microarray Data Sets
- Brain Tumor
- Lung Tumor

*http://algorithmics.molgen.mpg.de/Static/Supplements/-CompCancer/datasets.htm

As a pre-processing, Normalize the dataset along their column.
Data Preparation

Artificial Data Sets
- Arda25_30_3
- Arda50_75_5

Real Life Microarray Data Sets
- Brain Tumor
- Lung Tumor

*http://algorithmics.molgen.mpg.de/Static/Supplements/-CompCancer/datasets.htm

As a pre-processing, Normalize the dataset along their column.
Parameter Settings

Parameters

1. 20 trial runs
2. 100 generations under each run
3. 0.8 crossover rate
4. 0.01 mutation rate
5. 50 population size
6. Chromosome Length = Sum of genes and product of genes with cluster number.
Statistical Measures

CP Index

\[
CP(C, c) = \frac{s + d}{t} \times 100
\]  

(11)

Where \( s \) be the number of pairs belong to the same clusters in \( C \) (actual) and \( c \) (result), and \( d \) be the number of pairs belong to different clusters, and \( t \) be the total number of pairs in the data set, i.e., \( n_{c2} \).
## Statistical Measures

### R Index

<table>
<thead>
<tr>
<th>Actual Ranking</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking Solution 1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Ranking Solution 2</td>
<td>9</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cut 1</th>
<th>Cut 2</th>
<th>Cut 3</th>
<th>Cut 4</th>
<th>Cut 5</th>
<th>Cut 6</th>
<th>Cut 7</th>
<th>Cut 8</th>
<th>Cut 9</th>
<th>Cut 10</th>
<th>Cut 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/3</td>
<td>1/2</td>
<td>3/5</td>
<td>2/3</td>
<td>5/7</td>
<td>3/4</td>
<td>3/5</td>
<td>9/10</td>
<td>10/11</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>3/7</td>
<td>1/3</td>
<td>2/5</td>
<td>3/5</td>
<td>4/5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Statistical Measures

R Plot
Competitive Methods

- SOSCFRA_DB
- SOSCFRA_XB
- SOSCFRA_DX
- K-means
- FCM
- Hierarchical single linkage (HICSIL)
- Hierarchical average linkage (HICAL)
- Hierarchical complete linkage (HICCOL)
- Hierarchical centroid linkage (HICCEL)
- Hierarchical ward linkage (HICWAL)
Competitive Methods

- SOSCFRA_DB
- SOSCFRA_XB
- SOSCFRA_DX
- K-means
- FCM
  - Hierarchical single linkage (HICSIL)
  - Hierarchical average linkage (HICAL)
  - Hierarchical complete linkage (HICCOL)
  - Hierarchical centroid linkage (HICCEL)
  - Hierarchical ward linkage (HICWAL)
Competitive Methods

- SOSCFRA_DB
- SOSCFRA_XB
- SOSCFRA_DX
- K-means
- FCM
- Hierarchical single linkage (HICSIL)
- Hierarchical average linkage (HICAL)
- Hierarchical complete linkage (HICCOL)
- Hierarchical centroid linkage (HICCEL)
- Hierarchical ward linkage (HICWAL)
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline

1 Introduction

2 Multi-objective Optimization (MOO)
   - Definition and need of MOO
   - NSGA-II

3 Proposed Algorithm & its Components

4 Experiments and Results
   - Experimental Designs
   - Results and Discussion

5 Conclusions & Feature Scopes
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline

Introduction

Multi-objective Optimization (MOO)

Definition and need of MOO

NSGA-II

Proposed Algorithm & its Components

Proposed Algorithm & its Components

Experiments and Results

Experimental Designs

Results and Discussion

Conclusions & Feature Scopes

For Artificial Data Sets

Results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Arda25..30..3</th>
<th>Arda50..75..5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSCFRA</td>
<td>100.00(±0.000)</td>
<td>86.6036(±3.428)</td>
</tr>
<tr>
<td>SOSCFRA_DX</td>
<td>31.0345(±0.000)</td>
<td>18.9189(±0.000)</td>
</tr>
<tr>
<td>SOSCFRA_DB</td>
<td>31.0315(±0.000)</td>
<td>18.9189(±0.000)</td>
</tr>
<tr>
<td>SOSCFRA_XB</td>
<td>76.5977(±0.925)</td>
<td>84.8649(±0.680)</td>
</tr>
<tr>
<td>K-means</td>
<td>96.0115(±9.800)</td>
<td>84.3532(±2.371)</td>
</tr>
<tr>
<td>FCM</td>
<td>95.6322(±0.000)</td>
<td>80.5045(±0.388)</td>
</tr>
<tr>
<td>HICSIL</td>
<td>35.8621(±0.000)</td>
<td>25.3333(±0.000)</td>
</tr>
<tr>
<td>HICCOL</td>
<td>71.7241(±0.000)</td>
<td>81.6577(±0.000)</td>
</tr>
<tr>
<td>HICAL</td>
<td>91.7241(±0.000)</td>
<td>81.5495(±0.000)</td>
</tr>
<tr>
<td>HICCEL</td>
<td>77.2414(±0.000)</td>
<td>66.8468(±0.000)</td>
</tr>
<tr>
<td>HICWAL</td>
<td>91.7241(±0.000)</td>
<td>85.2973(±0.000)</td>
</tr>
</tbody>
</table>

Table: Experimental Result on Artificial Data Sets.

Discussion

Higher value of CP index and lower value of standard deviation in all artificial data set indicates that each time MOSCFRA identify clusters more correctly then other algorithms.
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline
Introduction
Multi-objective Optimization (MOO)
Definition and need of MOO
NSGA-II
Proposed Algorithm & its Components
Experiments and Results
Experimental Designs
Results and Discussion
Conclusions & Feature Scopes

For Artificial Data Sets

Results

For *Arda25_30_3* dataset.

Discussion

R plot shows the R-Index for the highest CP value of all runs in each algorithm (MOSCFRA, *SOSCFRA_DX*, *SOSCFRA_DB*, *SOSCFRA_XB*).
For Artificial Data Sets

Results

For Arda50_75_5 dataset.

Discussion

R plot shows the R-Index for the highest CP value of all runs in each algorithm (MOSCFRA, SOSCFRA_DX, SOSCFRA_DB, SOSCFRA_XB).
Discussion

These R Index or R Plot is only possible for the artificial datasets.

In Graph, R curve of MOSCFRA algorithm for both the datasets is near to 1 and on and average above all other curve. So, we can say, MOSCFRA generates the better ranking result.
For Artificial Data Sets

Discussion

These R Index or R Plot is only possible for the artificial datasets.

In Graph, R curve of MOSCFRA algorithm for both the datasets is near to 1 and on and average above all other curve. So, we can say, MOSCFRA generates the better ranking result.
For Real Life Data Sets

Results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Brain Tumor</th>
<th>Lung Tumor</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSCFRA</td>
<td>82.0209(± 8.515)</td>
<td>78.4193(± 3.618)</td>
</tr>
<tr>
<td>SOSCFRA_DX</td>
<td>19.6283(± 0.000)</td>
<td>49.4659(± 0.000)</td>
</tr>
<tr>
<td>SOSCFRA_DB</td>
<td>19.6283(± 0.000)</td>
<td>49.4659(± 0.000)</td>
</tr>
<tr>
<td>SOSCFRA_XB</td>
<td>81.9698(± 0.878)</td>
<td>76.7605(± 0.555)</td>
</tr>
<tr>
<td>K-means</td>
<td>73.8850(± 11.458)</td>
<td>65.5229(± 6.531)</td>
</tr>
<tr>
<td>FCM</td>
<td>69.1347(± 4.047)</td>
<td>60.4251(± 4.052)</td>
</tr>
<tr>
<td>HICSIL</td>
<td>30.3136(± 0.000)</td>
<td>56.5381(± 0.000)</td>
</tr>
<tr>
<td>HICCOL</td>
<td>44.0186(± 0.000)</td>
<td>71.6919(± 0.000)</td>
</tr>
<tr>
<td>HICAL</td>
<td>30.3136(± 0.000)</td>
<td>57.4111(± 0.000)</td>
</tr>
<tr>
<td>HICCEL</td>
<td>30.3136(± 0.000)</td>
<td>57.4111(± 0.000)</td>
</tr>
<tr>
<td>HICWAL</td>
<td>67.0151(± 0.000)</td>
<td>77.1237(± 0.000)</td>
</tr>
</tbody>
</table>

Table: Experimental Result on Real Life Data Sets.

Discussion

Higher value of CP index and lower value of standard deviation in all real data set indicates that each time MOSCFRA identify clusters more correctly then other algorithms.
Results:

Top 10 genes in Brain Tumor Data Set
'S81957_at', 'D38500_at', 'K02268_at', 'X64072_s_at', 'M58297_at', 'J04132_at', 'M93119_at', 'J04444_at', 'L36847_at', 'HG3141-HT3317_f_at'

Top 10 genes from Lung Tumor Data Set
'39022_at', '939_at', '32251_at', '33373_at', '37849_at', '40195_at', '32034_at', '40647_at', '33273_f_at', '34335_at'
Results:

Top 10 genes in Brain Tumor Data Set
'S81957_at', 'D38500_at', 'K02268_at', 'X64072_s_at',
'M58297_at', 'J04132_at', 'M93119_at', 'J04444_at',
'L36847_at', 'HG3141-HT3317_f_at'

Top 10 genes from Lung Tumor Data Set
'39022_at', '939_at', '32251_at', '33373_at', '37849_at',
'40195_at', '32034_at', '40647_at', '33273_f_at', '34335_at'
Discussion

In every data sets, generated CP index by the \textit{SOSCFRA\_DX} is very small. In case of \textit{SOSCFRA\_XB}, the CP index is some times slightly greater than our proposed multiobjective part, MOSCFRA.

When the DB index and XB index merge into our single objective counter part, \textit{SOSCFRA\_DX}, it gives the same result as given in the \textit{SOSCFRA\_DB}.

Since DB index is affecting the goodness of XB index. We can say, DB index is not a good index in this case.
In every data sets, generated CP index by the \textit{SOSCFRA\_DX} is very small. In case of \textit{SOSCFRA\_XB}, the CP index is some times slightly greater than our proposed multiobjective part, MOSCFRA.

When the DB index and XB index merge into our single objective counter part, \textit{SOSCFRA\_DX}, it gives the same result as given in the \textit{SOSCFRA\_DB}.

Since DB index is affecting the goodness of XB index. We can say, DB index is not a good index in this case.
Discussion

In every data sets, generated CP index by the \textit{SOSCFRA\_DX} is very small. In case of \textit{SOSCFRA\_XB}, the CP index is some times slightly greater than our proposed multiobjective part, MOSCFRA.

When the DB index and XB index merge into our single objective counter part, \textit{SOSCFRA\_DX}, it gives the same result as given in the \textit{SOSCFRA\_DB}.

Since DB index is affecting the goodness of XB index. We can say, DB index is not a good index in this case.
Simultaneous Clustering and Gene Ranking

Kartick Chandra Mondal

Outline

1. Introduction

2. Multi-objective Optimization (MOO)
   - Definition and need of MOO
   - NSGA-II

3. Proposed Algorithm & its Components
   - Proposed Algorithm & its Components

4. Experiments and Results
   - Experimental Designs
   - Results and Discussion

5. Conclusions & Feature Scopes
   - Conclusions & Feature Scopes
Conclusions

Address the problem of unsupervised gene ranking and clustering.

Proposed an algorithm in clustering and feature ranking research area which use multiobjective framework (NSGA-II) for simultaneous operation.

An new encoding technique is developed for our problem and the optimization criteria, XB index and DB index, are minimized simultaneously.

Performance is measured by using two artificial data sets and two real life cancer data sets.
Conclusions

Address the problem of unsupervised gene ranking and clustering.

Proposed an algorithm in clustering and feature ranking research area which use multiobjective framework (NSGA-II) for simultaneous operation.

An new encoding technique is developed for our problem and the optimization criteria, XB index and DB index, are minimized simultaneously.

Performance is measured by using two artificial data sets and two real life cancer data sets.
Conclusions

Address the problem of unsupervised gene ranking and clustering.

Proposed an algorithm in clustering and feature ranking research area which use multiobjective framework (NSGA-II) for simultaneous operation.

An new encoding technique is developed for our problem and the optimization criteria, XB index and DB index, are minimized simultaneously.

Performance is measured by using two artificial data sets and two real life cancer data sets.
Conclusions

Address the problem of unsupervised gene ranking and clustering.

Proposed an algorithm in clustering and feature ranking research area which use multiobjective framework (NSGA-II) for simultaneous operation.

An new encoding technique is developed for our problem and the optimization criteria, XB index and DB index, are minimized simultaneously.

Performance is measured by using two artificial data sets and two real life cancer data sets.
Future Scopes

Try to implement the algorithm for unknown number of clusters and on other multiobjective algorithms other than NSGA-II.

Use different statistical measures to find the algorithmic performance.

Need a closer look at choosing objective function and during Pareto optimal set selection.
Future Scopes

Try to implement the algorithm for unknown number of clusters and on other multiobjective algorithms other than NSGA-II.

Use different statistical measures to find the algorithmic performance.

Need a closer look at choosing objective function and during Pareto optimal set selection.
Future Scopes

Try to implement the algorithm for unknown number of clusters and on other multiobjective algorithms other than NSGA-II.

Use different statistical measures to find the algorithmic performance.

Need a closer look at choosing objective function and during Pareto optimal set selection.
Thank you.