Metaheuristic Optimization and Clustering

Metaheuristic Optimization

- For solving a very general class of computational problems
- Combining user-given black-box procedures
- Applied to problems for which there is no satisfactory problem-specific algorithm or heuristic/not practical to apply such algorithms
- Generally targeted to combinatorial optimization problems

General Flow of Control

- State/potential solution
  - Single or a pool of solutions
- An objective function to be optimized
- Procedure for generating new solutions probabilistically
- Probabilistic acceptance of one/more new solutions
- Iterative procedure
- Examples: Genetic algorithms, simulated annealing, evolutionary strategies, tabu search, ant colony optimization, artificial immune systems

GENETIC ALGORITHMS

- Definition

  Randomized search and optimization technique guided by the principles of natural genetic systems.

- Why Genetic Algorithms (GAs) ?
  - Evolution produced good individuals, similar principles might work for solving complex problems
  - Many problems can not be solved in polynomial amount of time using a deterministic algorithm
  - Near optimal solutions requiring less time more desirable than optimal solutions with huge amount of time
    - E.g., traveling salesman problem, knapsack problem

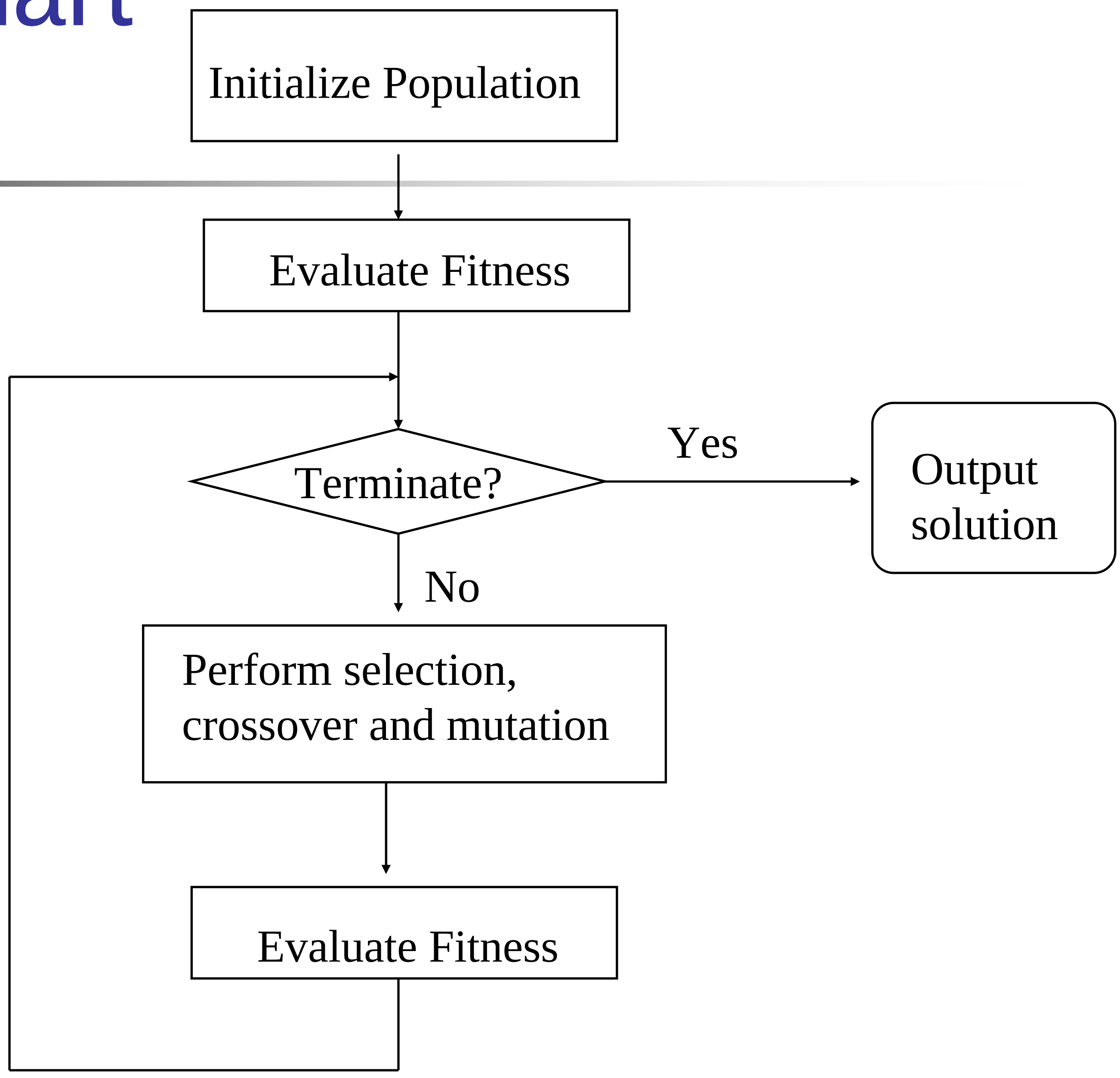
Genetic Algorithms - Features

- Evolutionary Search and Optimization Technique
- Principles of Evolution (*survival of the fittest* and *inheritance*)
- Work with coding of the parameter set
- Searches from a population of points
- Uses probabilistic transition rules

Genetic Algorithms ⟺ Nature

Genetic Algorithms	Nature
•A solution (phenotype)	Individual
•Representation of a solution (genotype)	Chromosome
•Components of the solution	Genes
•Set of solutions	Population
•Survival of the fittest (selection)	Darwins theory
•Search operators	Crossover and mutation
• Iterative procedure	Generations

GA Flowchart



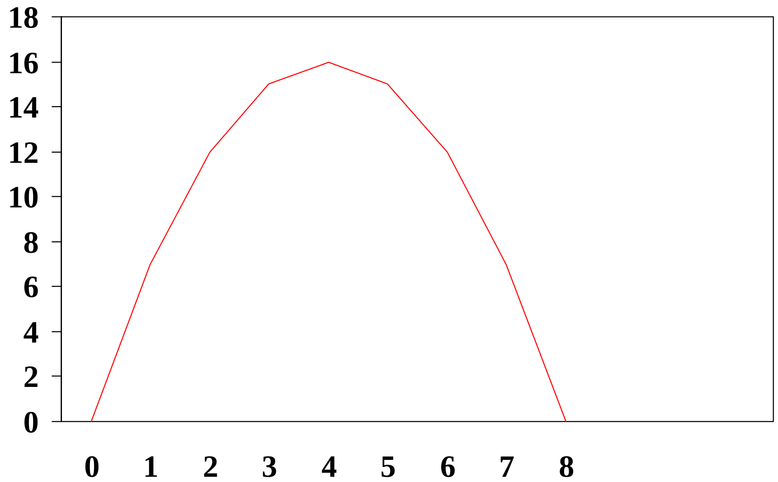
Encoding Strategy and Population

- Chromosome encodes a solution in the search space
  - Usually as strings of 0's and 1's
  - If *l* is the string length, number of different chromosomes (or strings) is 2^l
- Population
  - A set of chromosomes in a generation
  - Population size is usually constant
  - Common practice is to choose the initial population randomly.

Encoding and Population - Example

Optimization Problem :

Optimize $f(x) = x(8 - x)$, $x=[0,8]$



User specified parameters

Binary String of 8 bits

0-255 ⟷ 0-8

Chromosome encodes *x*

1	**0**	**0**	**1**	**1**	**0**	**1**	**0**

Value = 154 ⟶ 8/255 * 154 + 0 = 4.8313

Encoding and Population – Example contd...

Population (size = 4)	**Corresponding x**
1 0 0 1 1 0 1 0	4.8313
0 1 1 0 0 1 1 1	3.2313
0 0 0 1 0 1 0 1	0.6588
1 0 1 1 1 1 0 0	5.8980

Fitness Evaluation

- Fitness/objective function associated with each chromosome
- indicates the degree of goodness of the encoded solution
- only problem specific information (also known as the payoff information) that GAs use
- If minimization problem is to be solved then **fitness $\propto$ 1/objective**.

Fitness Evaluation - Example

Function *f(x) = x(8-x)*

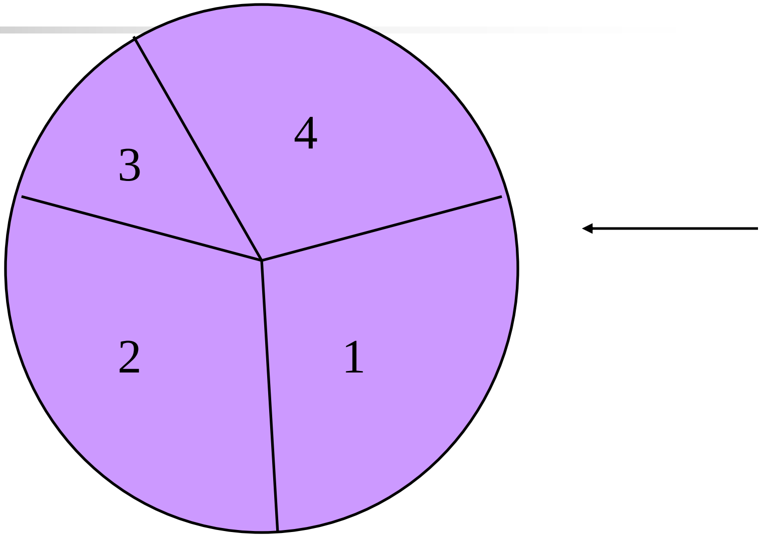
Population (size = 4)	**Corresponding x**	**Objective/ Fitness fn.**
1 0 0 1 1 0 1 0	4.8313	15.3089
0 1 1 0 0 1 1 1	3.2313	15.4091
0 0 0 1 0 1 0 1	0.6588	4.8363
1 0 1 1 1 1 0 0	5.898	12.3975

Selection

- More copies to good strings
- Fewer copies to bad string
- proportional selection scheme
  - Number of copies taken to be directly proportional to its fitness
  - mimics the natural selection procedure to some extent.
  - Roulette wheel parent selection and stochastic universal selection selection are two frequently used selection procedures.

Roulette Wheel Selection – Example

Chromosome #	Fitness
1	15.3089
2	15.4091
3	4.8363
4	12.3975

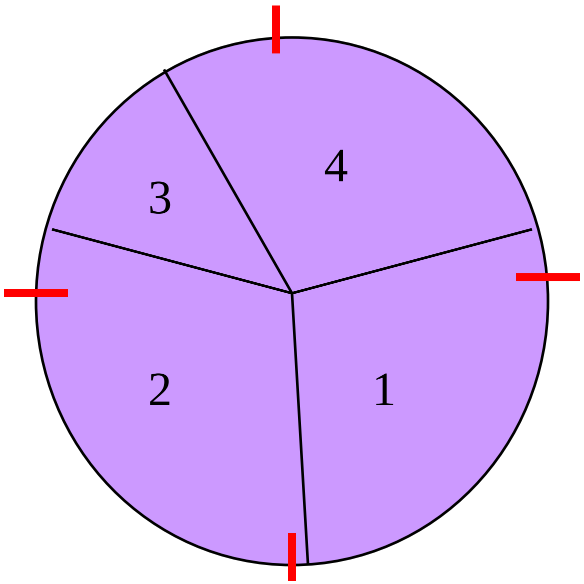


Spin 1	Chromosome 2 selected
Spin 2	Chromosome 1 selected
Spin 3	Chromosome 2 selected
Spin 4	Chromosome 4 selected

Mating Pool

0 1 1 0 0 1 1 1
1 0 0 1 1 0 1 0
0 1 1 0 0 1 1 1
1 0 1 1 1 1 0 0

Stochastic Universal Selection- Example



Chromosome 1	1 copy
Chromosome 2	2 copies
Chromosome 3	0 copies
Chromosome 4	1 copy

Crossover

- **exchange information**
  - between randomly selected parent chromosomes
  - Single point crossover is one of the most commonly used schemes.
  - probabilistic operation

Crossover – Example

1	**0**	**0**	**1**	**1**	**0**	**1**	**0**

0	**1**	**1**	**1**	**1**	**1**	**0**	**1**

Here *l* (string length) = 8. Let *k* (crossover point) = 5

Offspring formed :

1	**0**	**0**	**1**	**1**	**1**	**0**	**1**

0	**1**	**1**	**1**	**1**	**0**	**1**	**0**

Mutation

- **random alteration** in the genetic structure
  - introduce **genetic diversity** into the population.
  - Exploration of new search areas
  - Mutating a binary gene involves simple **negation of the bit**,
  - Mutating a real coded gene defined in a variety of ways
  - probabilistic operation

Mutation- Example

1	**0**	**0**	**1**	**1**	**0**	**1**	**0**

Mutations at positions 2 and 5

1	**1**	**0**	**1**	**0**	**0**	**1**	**0**

Termination Criterion

The cycle of selection, crossover and mutation is repeated a number of times till:

• the average fitness value of a population becomes more or less constant over a specified number of generations,

• a desired objective function value is attained by at least one string in the population,

• the number of generations (or iterations) is greater than some threshold ----- most commonly used.

Elitist Model of GAs

Best string seen up to the current generation is preserved

Parameters ...

- population size (usually fixed)
- string length (usually fixed)
- probabilities of performing crossover (μ_c) and mutation(μ_m),
  - μ_c is kept high and μ_m is kept low
- the termination criteria
  - Generally a maximum number of iterations
- parameters are user determined and problem dependent
- no firm guidelines
- parameters can be kept variable and/or adaptive.

Parameters – Example

For the example being considered,

$P = 4$, $l = 8$.

But for most realistic cases

P is usually chosen in the range 50-100.

$\mu_c = [0.6\text{-}0.9]$,

$\mu_m = [0.01\text{-}0.1]$.

l usually depends on the required precision

Current Trends in GAs

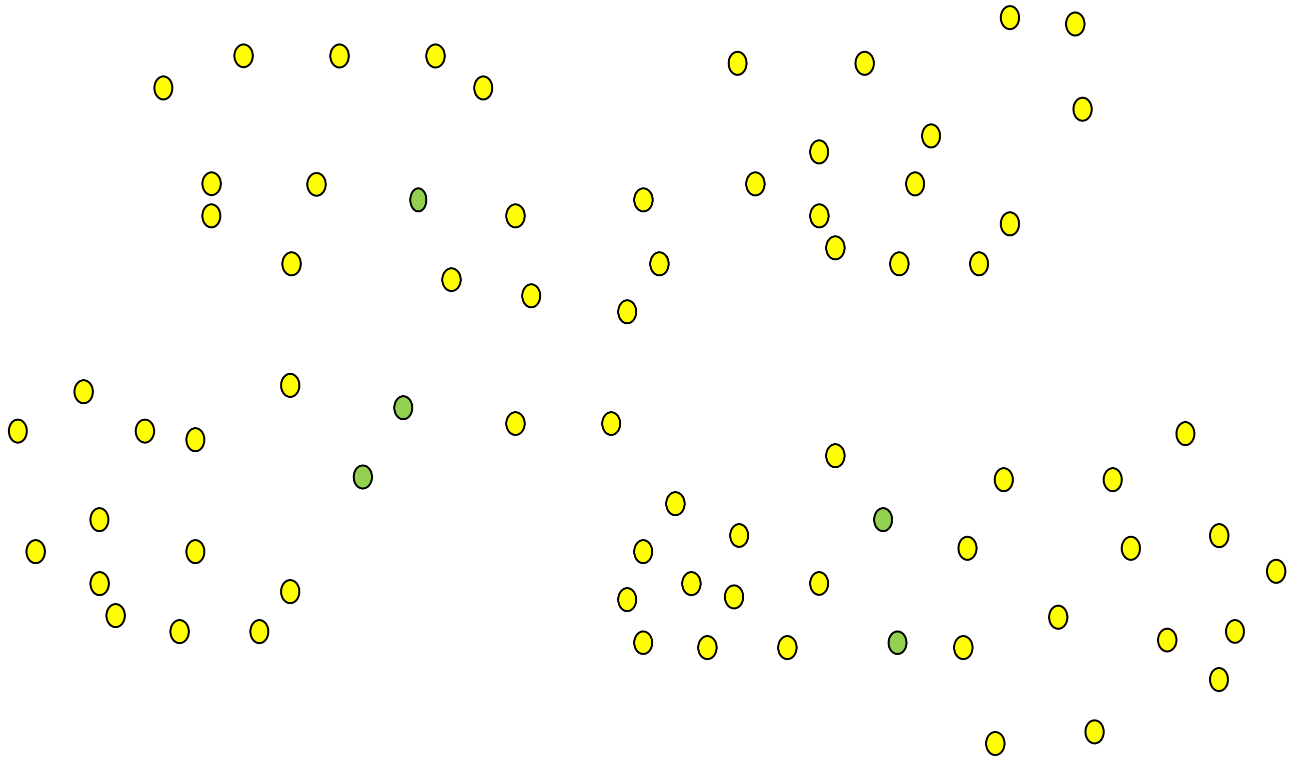
- Encoding strategy
  - Integer encoding
  - Real encoding
  - Encoding of other structures
  - Variable length representation
- Operators
  - New domain specific operators
  - Variable and adaptive probabilities of the operations
- Incorporation of local search
- Handling constraints
  - Reject infeasible strings
  - Penalty based method

Current Trends in GAs contd...

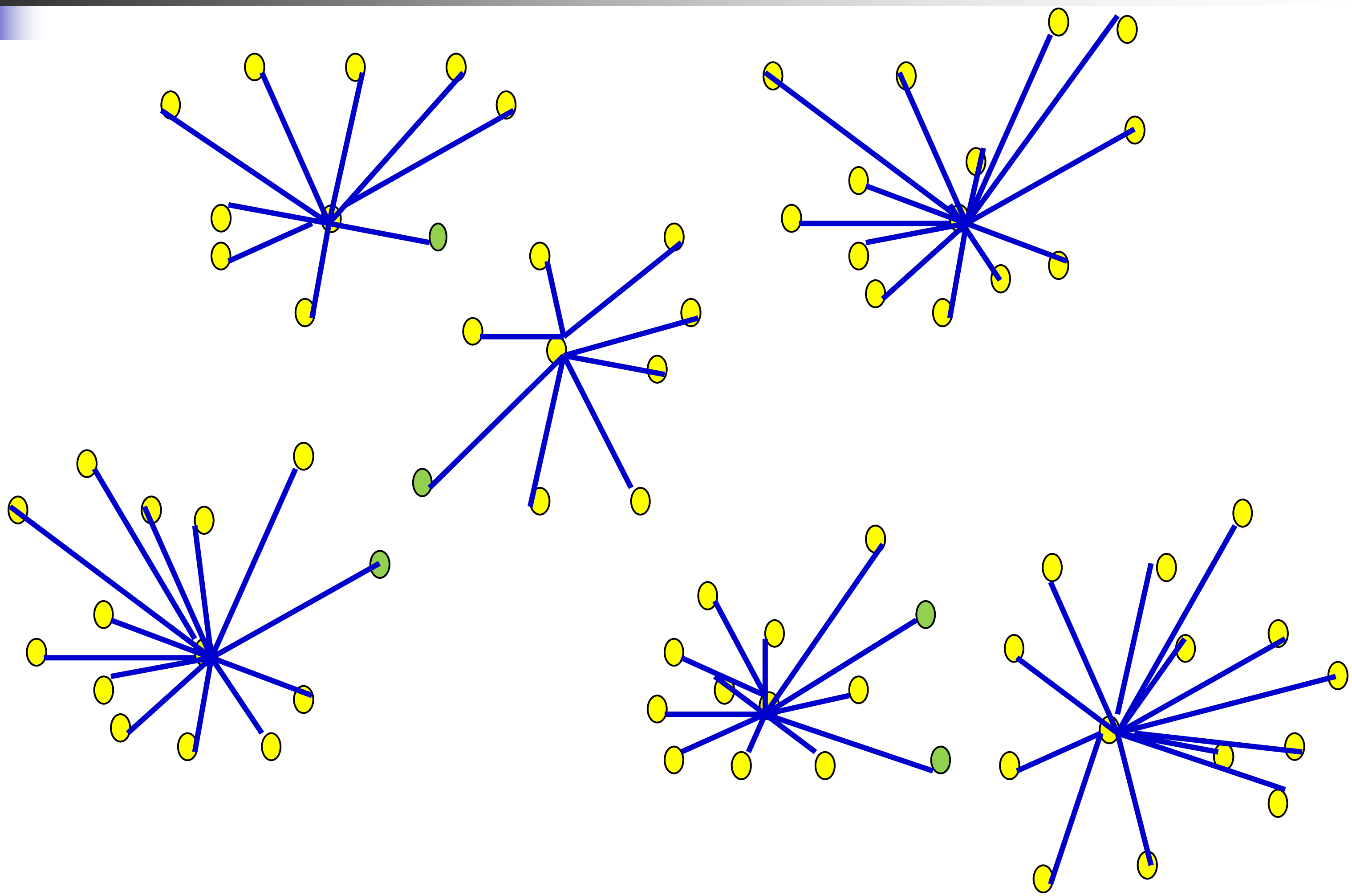
- Multiobjective optimization
  - Multiple conflicting objectives to be simultaneously optimized
  - NSGA-II, PAES, SPEA, AMOSA
- Hybridization with other soft computing tools like
  - Neural networks
  - Fuzzy sets
  - Rough sets

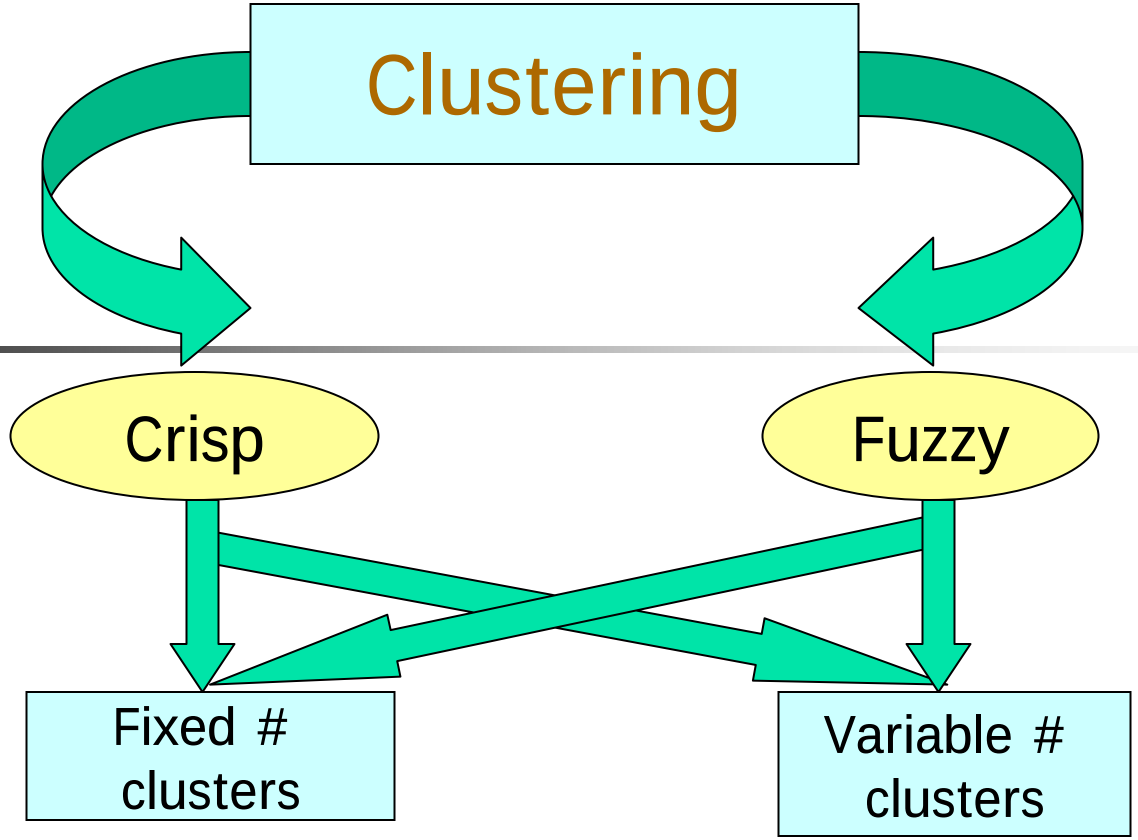
Genetic Algorithms for Clustering

Spatial Clustering



Spatial Clustering





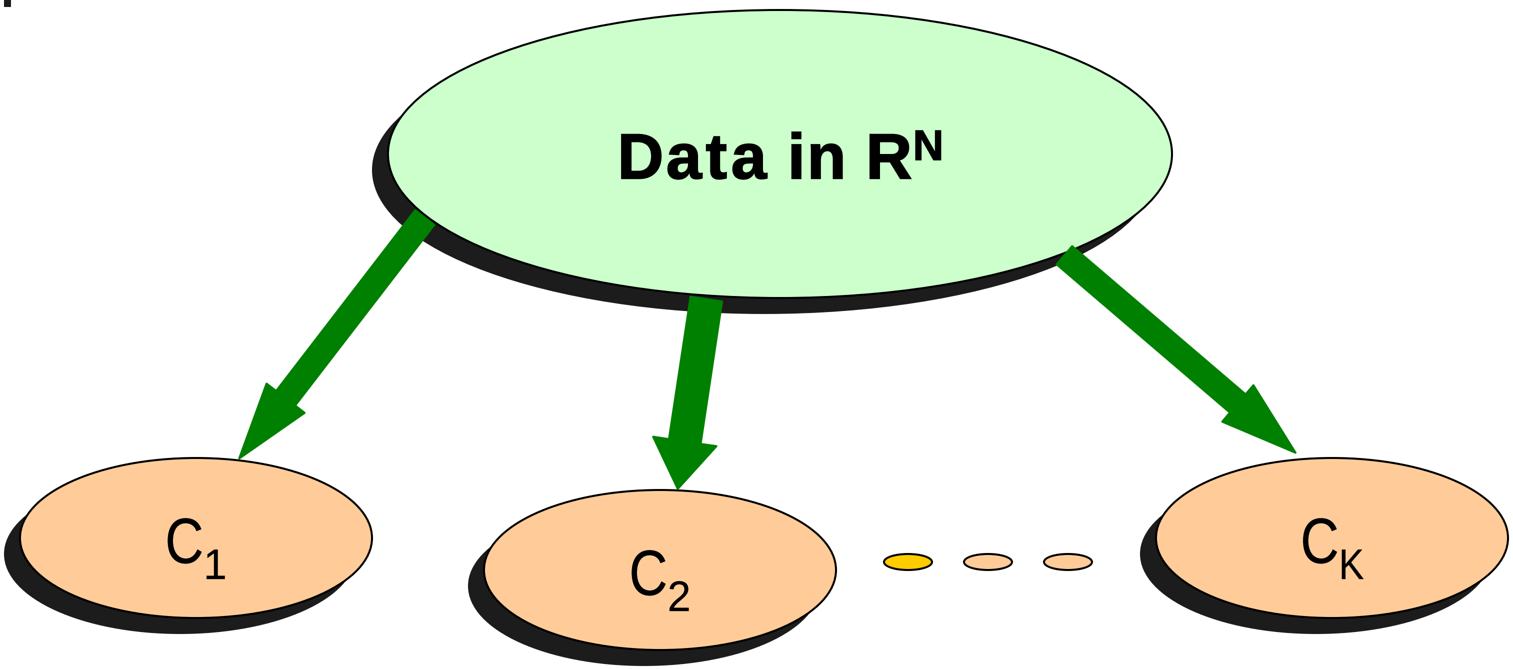
"Genetic Algorithm Based Clustering Technique'', Patt. Recog., 33, 1455-1465, 2000.

"Clustering using Simulated Annealing with Probabilistic Redistribution'' Int. Journal of Pattern Recognition and Artificial Intelligence, vol 15, no. 2, pp. 269-285, 2001.

"Non-parametric Genetic Clustering : Comparison of Validity Indices'', IEEE Trans. on Systems, Man and Cybernetics Part-C, vol. 31, no. 1, pp. 120-125, 2001.

''Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification'', Pattern Recognition.

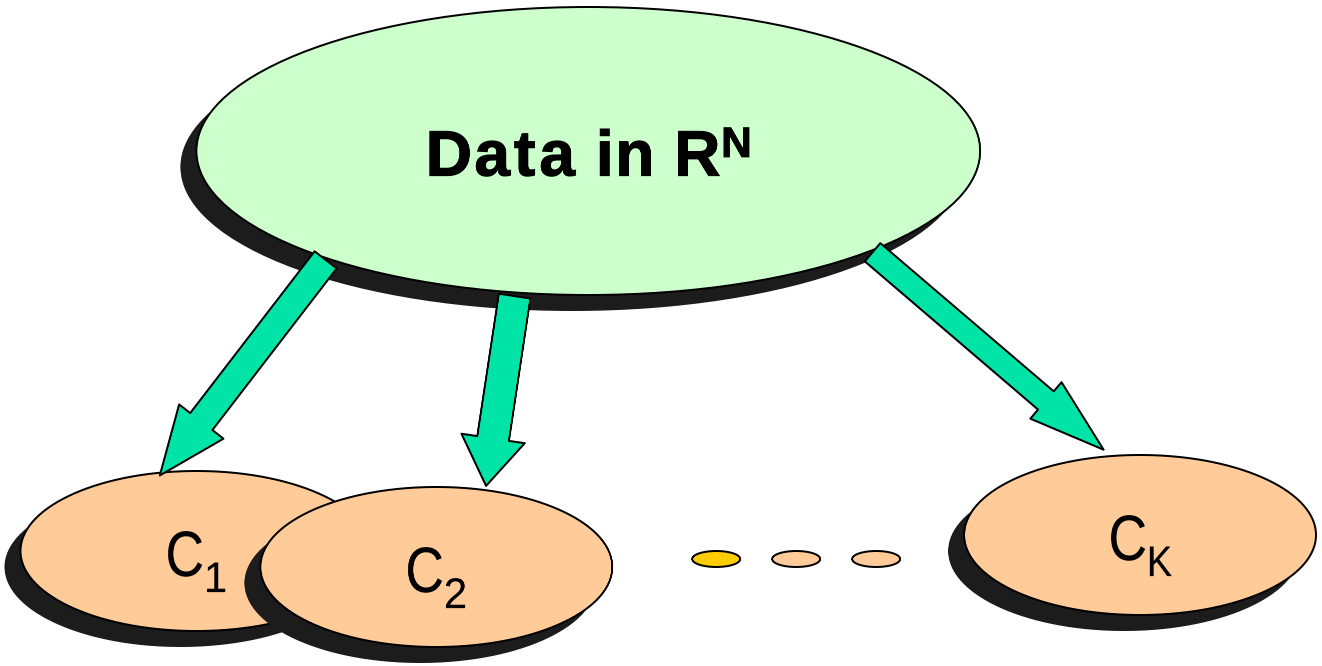
Crisp Clustering



$C\_i \neq Ø$, $C\_i \cap C\_j = Ø$

Some measure of clustering goodness is optimized

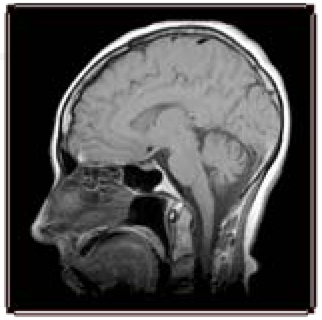
Fuzzy Clustering



$C\_i \neq Ø$, $C\_i \cap C\_j \neq Ø$

Some measure of clustering goodness is optimized

Applications of Clustering



- Medical Image Segmentation
  - MRI brain image
  - X-Ray Computer Tomography (CT)
- Spatial Data Mining
  - Creating thematic maps in GIS
- WWW
  - Clustering the documents
- Market economics
  - Financial analysis
    - Clustering of time series

Important Issues in Clustering

- Data types, for interval-scaled, boolean, nominal, ordinal and ratio variables.
  - Ordinal variables – Rank in a class
  - Nominal – categorical (red, blue, green)
  - Interval scaled – continuous variables
- Distance/similarity measures
  - Minkowski distance $\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\ldots+|x_{ip}-x_{jp}|^q)}$
    - q=1 →Manhattan, q=2→ Euclidean, Pearson correlation, Mahalanobis distance, symmetry-based, binary/categorical
- Cluster types in the data
  - Model selection
  - Hyperspherical, elliptical, ring like, arbitrary shaped
  - Choice of algorithm and distance measure

Important Issues in Clustering
contd...

- How many clusters
  - Model order selection
- Clustering quality/goodness
  - Depends on similarity measure
  - Optimizing criteria
    - Low intra class variance
    - High inter class variance
    - Cluster Validity
  - Optimizing technique
    - Gradient descent
    - Meta heuristic approaches

Methods

- K-means clustering
- Fuzzy c-means clustering
- Hierarchical clustering

Clustering Using Genetic Algorithms (fixed *c*)

- ***Representation :***
  - Cluster centers encoded in the chromosomes

    For a *d*-dimensional space

    length of chromosome = $d * K$

$$\{ \underbrace{(v_{11}, v_{12}, \ldots, v_{1d})}_{\text{Center 1}} \underbrace{(v_{21}, v_{22}, \ldots, v_{2d})}_{\text{Center 2}} \ldots \underbrace{(v_{c1}, v_{c2}, \ldots, v_{cd})}_{\text{Center } c} \}$$

- **Example 1 →**

  Let $d=2$ and $K=3$,
  - i.e., two-dimensional space, number of clusters = 3

    Chromosome ➔ {(51.6 72.3) (18.3 15.7) (29.1 32.2)}

    represents 3 cluster centers

    (51.6, 72.3), (18.3, 15.7) and (29.1, 32.2).

GAs for Clustering

(Population initialization)

- Initial cluster centers = *c* randomly selected points from the data

```
For each chromosome i in the population
      For each cluster j
                p=randomly chosen point from the data set;
                Population[i][j] ← p;
      End
End
```

GAs for Clustering
(*Fitness computation*)

This consists of three phases.

- **Phase 1**: Cluster assignment
  - each point is assigned to the nearest cluster center.

    All ties are resolved arbitrarily.

- **Phase 2**: The cluster centers encoded in the chromosome are replaced by the mean points of the respective clusters.
  - $v_i^* = (1/n_i)\, \Sigma\, x_j^i$ for $i=1, 2, \ldots, c$
  - v_i replaced by v_i^* in the chromosome

- **Phase 3:** fitness = 1/(clustering metric J)
  - Compute $\boldsymbol{J = \Sigma\, \Sigma\, d^2(x_k^j, v_j)}$, $j=1,2,\ldots c$ and $k=1, 2, \ldots, n_j$.
  - Maximization of fitness leads to minimization of J

GAs for Clustering

(Fitness computation - Example)

Example 2 →

Chromosome ➔ {(51.6 72.3) (18.3 15.7) (29.1 32.2)}

- The first cluster center is (51.6, 72.3).

- Let points (50.0, 70.0) and (52.0, 74.0) be also included in the first cluster
  - besides itself i.e., (51.6, 72.3)
- Hence the newly computed cluster center becomes

  ((50.0+52.0+51.6)/3, (70.0+74.0+72.3)/3)=(51.2, 72.1).
- New cluster center replaces the previous value in chromosome

  (51.2, 72.1) replaces (51.6, 72.3).
- Compute mean squared error *J*
  - Fitness of chromosome = *1/J*

GAs for Clustering

(Genetic operations)

•***Crossover:***

Single point crossover with a fixed crossover probability.

•For chromosomes of length *c*

• a random integer *p* is generated in the range *[1, c]*

• portions of the chromosomes lying to the right of *p* are exchanged to produce two offspring.

•***Mutation:***

Since we are considering floating point representation, we use the following mutation. A number δ in the range [0, 1] is generated with uniform distribution. If the value at a gene position is v, after mutation it becomes

$$v=v \pm 2 * \delta * v, \text{ if } v \neq 0,$$

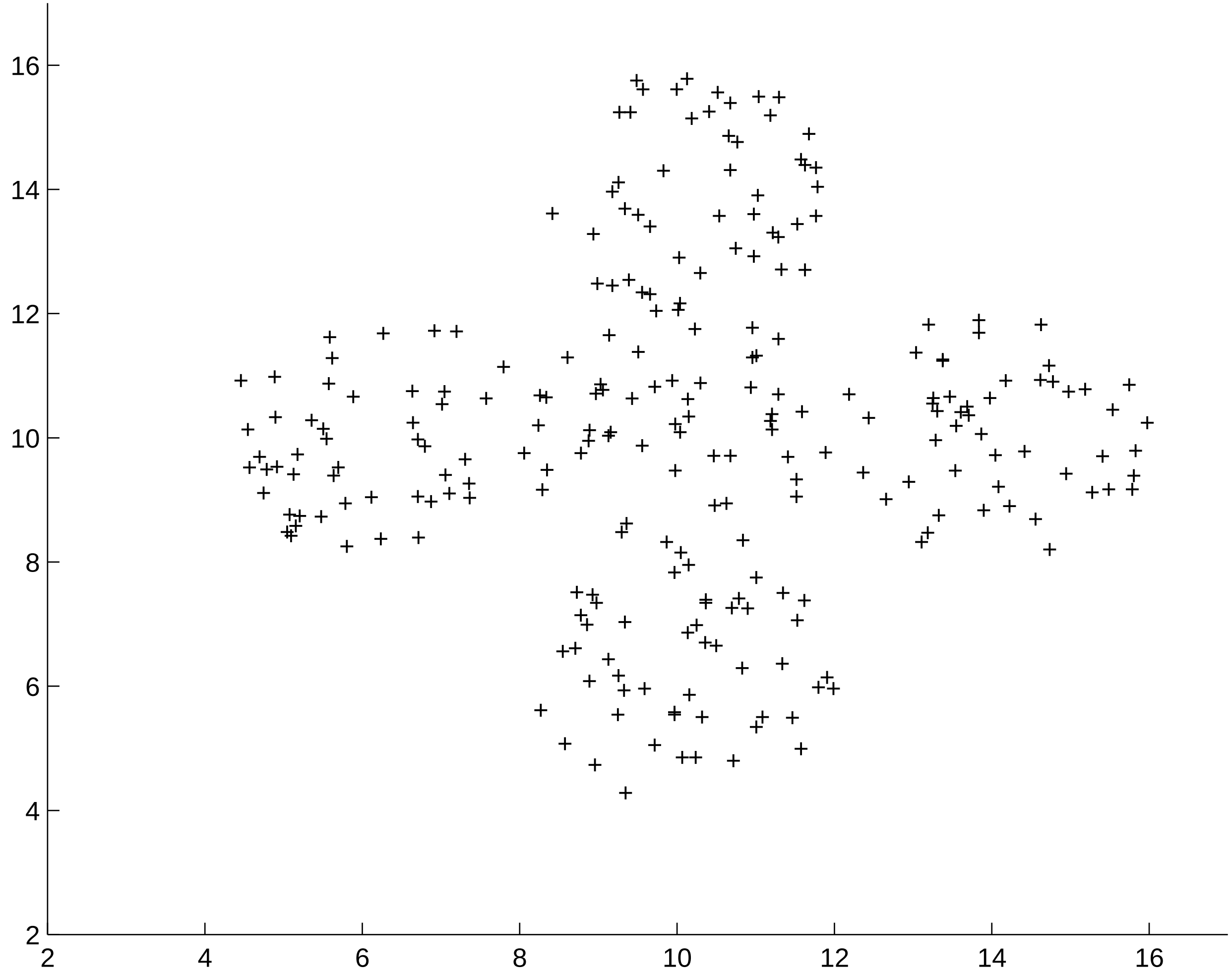
$$v=v \pm 2 * \delta \quad \text{if } v=0.$$

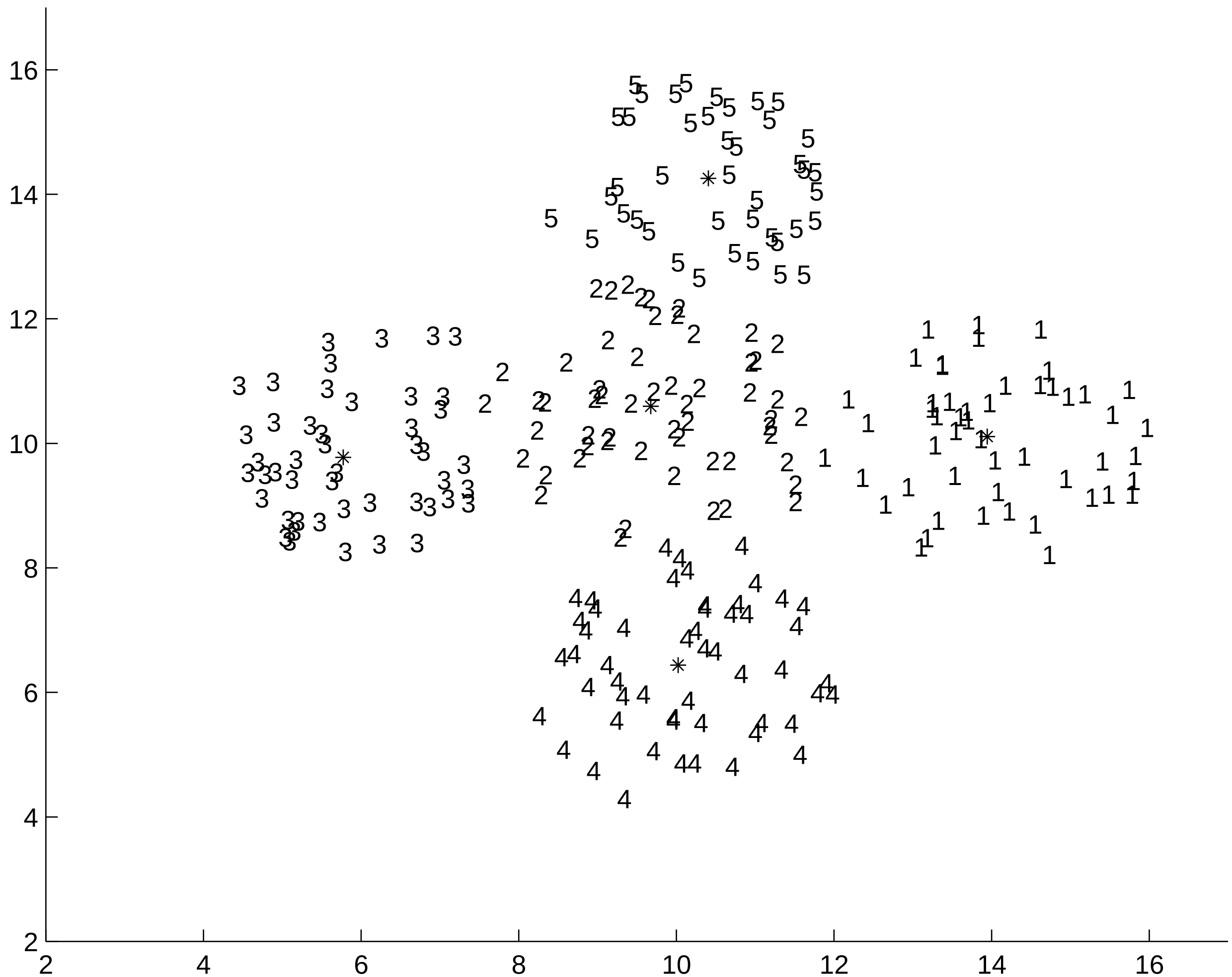
GAs for Clustering

- Termination
  - GA clustering is run for a fixed number of generations
  - Elitism incorporated
  - best string (one with the lowest J) is taken as the solution of the clustering problem.

Result

(c=5, n=250, d=2, iter=100, Pop = 20, $Prob_{crossover}$ =0.8, $Prob_{mutation}$ =0.01)




16
14
12
10
8
6
4
2
2 4 6 8 10 12 14 16

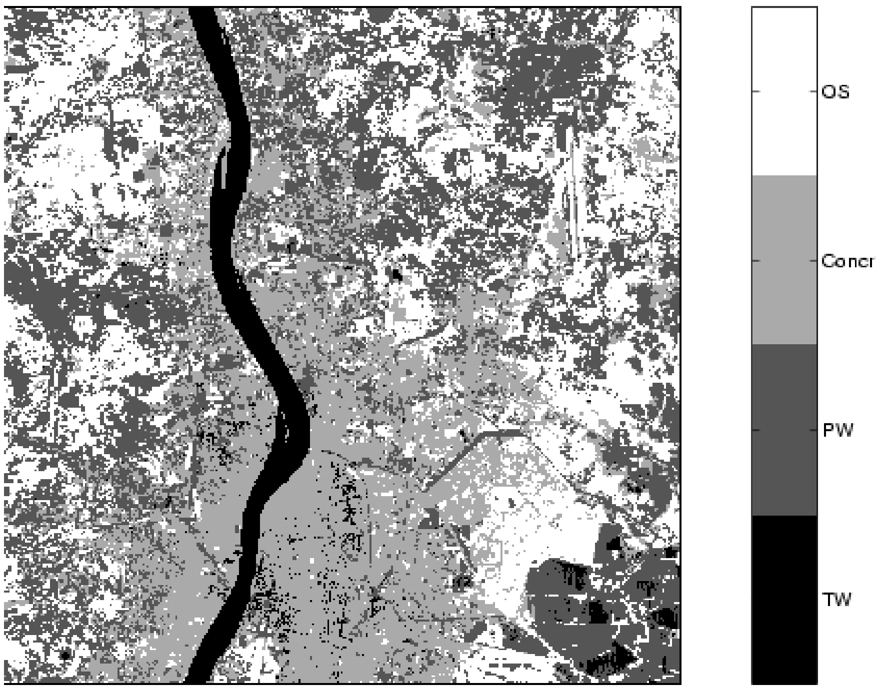
Pixel Classification of Satellite Images

- IRS image of Mumbai and Kolkata (IRS 1A)
- Four bands
  - Blue
  - Green
  - Red
  - Near infra red
- Resolution = 36.25 m x 36.25 m

Calcutta Image

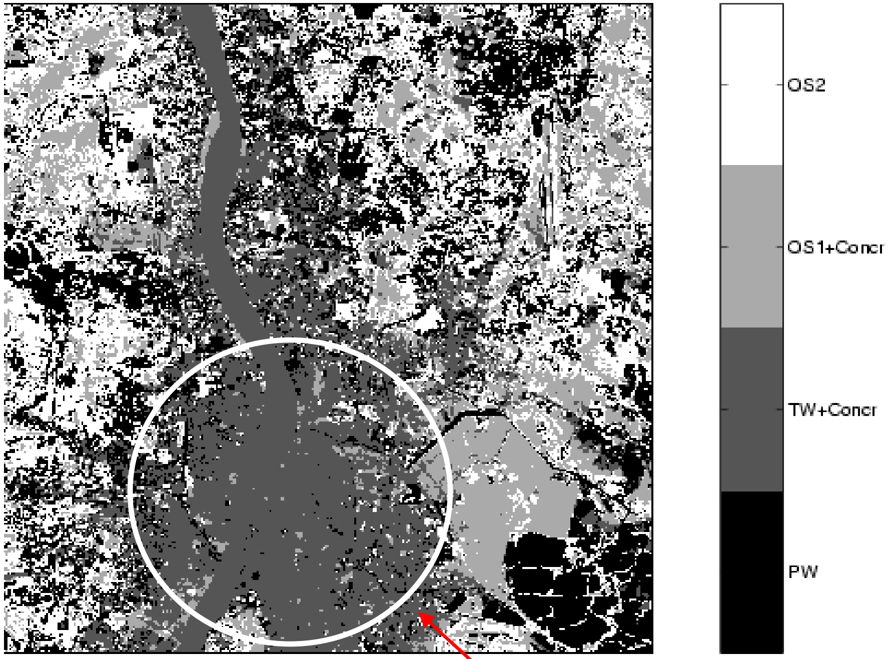
Input Image in Near Infra Red Band

Result on Calcutta Image



Genetic Fuzzy Clustering

clusters detected = 4



FCM Clustering for

clusters = 4

Confusion between water body and concrete

- *River, roads, fishery, Salt Lake region etc. automatically identified*
- *In FCM confusion between Water/open space and Concrete classes*
- *Index value for genetic scheme was better than that for FCM*

Multiobjective Clustering

Summary

- Unsupervised
- Many approaches
  - K-means – simple, sometimes useful
    - K-medoids is less sensitive to outliers
  - Hierarchical clustering – works for symbolic attributes
- Evaluation is a problem
  - Cluster validity

References

- J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, Reading, 1974.
- P. A. Devijver and J. Kittler, Pattern Recognition : A Statistical Approach, Prentice-Hall, London, 1982.
- K. Fukunaga, Introduction to Statistical Pattern Recognition (Second Edition), Academic Press, New York, 1990.
- S. Theodoridis, K. Koutroumbas, Pattern recognition, Academic Press, 1999.
- A. Webb, Statistical pattern recognition, Oxford University Press Inc., New York, 1999.
- R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification (2nd ed.), John Wiley and Sons, 2001.
- S. Bandyopadhyay, ``An Efficient Technique for Superfamily Classification of Amino Acid Sequences: Feature Extraction, Fuzzy Clustering and Prototype Selection", Fuzzy Sets and Systems, vol. 152, pp. 5-16, 2005
- J. T. L. Wang, Q. C. Ma, D. Shasha, C. H. Wu, ``New Techniques for Extracting Features from Protein Sequences”, IBM Systems Journal, Special Issue on Deep Computing for the Life Sciences, vol-40, no-2, pp. 426-441, 2001

References

- A. K. Jain, M.N. Murthy and P.J. Flynn, Data Clustering: A Review, ACM Computing Reviews, Nov 1999.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data, Prentice Hall, 1988.
- R. L. Cannon, J. V. Dave, and J. C. Bezdek. Efficient implementation of the fuzzy c-means clustering algorithms. TPAMI, 8(2):248--255, 1986.
- U. Maulik and S. Bandyopadhyay, ”Genetic Algorithm Based Clustering Technique“, Pattern Recognition, vol. 33, no. 9, pp. 1455-1465, 2000
- S. Bandyopadhyay and U. Maulik, “Non-parametric Genetic Clustering : Comparison of Validity Indices", *IEEE TSMC-C, vol. 31, no. 1, pp. 120-125, 2001*.
- S. Bandyopadhyay, “Simulated Annealing Using Reversible Jump Markov Chain Monte Carlo Algorithm for Fuzzy Clustering", IEEE TKDE, vol. 17, no. 4, pp. 479-490, 2005.

References

- S. Bandyopadhyay and S. K. Pal, Classification and Learning Using Genetic Algorithms, Applications in Bioinformatics and Web Intelligence, Springer-Verlag, 2007
- D. E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, New York, 1989.
- Michalewicz, Z. 1996. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag. Third edition.
- Mitchell, Melanie. 1996. *An Introduction to Genetic Algorithms*. Cambridge, MA
- Beyer, Hans-Georg. 2001. *The Theory of Evolution Strategies*. Heidelberg: Springer-Verlag.
- Schwefel, Hans-Paul. 1995. *Evolution and Optimum Seeking*. New York, NY: John Wiley.
- Fogel, David B. 1991. *System Identification through Simulated Evolution*. Needham Heights, MA: Ginn Press.
- Kalyanmoy Deb. Multi-Objective Optimization using Evolutionary Algorithms, John Wiley & Sons, Chichester, UK, 2001, ISBN 0-471-87339-X.
- Carlos A. Coello Coello, David A. Van Veldhuizen and Gary B. Lamont, Evolutionary Algorithms for Solving Multi-Objective Problems, Kluwer Academic Publishers, New York, March 2002, ISBN 0-3064-6762-3.

References

- S. Bandyopadhyay, S. K Pal, and B. Aruna, "Multi-objective GAs, quantitative Indices and Pattern Classification", *IEEE Transactions on Systems, Man and Cybernetics - B,* vol. 34, no. 5, pp. 2088-2099, 2004.
- U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 12, pp. 1650-1654, 2002.
- U. Maulik and S. Bandyopadhyay, "Fuzzy Partitioning Using Real Coded Variable Length Genetic Algorithm for Pixel Classification", *IEEE Transactions on Geosciences and Remote Sensing,* vol. 41, no. 5, pp. 1075-1081, 2003.
- S. Bandyopadhyay, "Simulated Annealing Using Reversible Jump Markov Chain Monte Carlo Algorithm for Fuzzy Clustering", *IEEE Transactions on Knowledge and Data Engineering,* vol. 17, no. 4, pp. 479-490, 2005.
- S. Bandyopadhyay, "An Efficient Technique for Superfamily Classification of Amino Acid Sequences: Feature Extraction, Fuzzy Clustering and Prototype Selection", *Fuzzy Sets and Systems,* vol. 152, pp. 5-16, 2005.
- S. Bandyopadhyay, Ac. Bagchi and U. Maulik, ``Active Site Driven Ligand Design: An Evolutionary Approach", *Journal of Bioinformatics and Computational Biology*, vol. 3, No. 5, pp.1053-1070, 2005.

References

- M. K. Pakhira, S. Bandyopadhyay and U. Maulik, ``A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification'', *Fuzzy Sets and Systems*, vol. 155, pp. 191-214, 2005.
- S. Bandyopadhyay, U. Maulik and A. Mukhopadhyay, ``Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery'', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1506-1511, 2007.
- S. S. Ray, S. Bandyopadhyay, and S. K. Pal, ``Genetic Operators for Combinatorial Optimization in TSP and Microarray Gene Ordering'', *Applied Intelligence*, vol. 26, no. 3, pp. 183-195, 2007.
- S. Bandyopadhyay and S. Saha, ``GAPS: A New Symmetry Based Genetic Clustering Technique", *Pattern Recognition,* vol. 10, no. 12, pp. 3430-3451, 2007.
- S. Bandyopadhyay, S. Saha, U. Maulik and K. Deb, ``A Simulated Annealing Based Multi-objective Optimization Algorithm: AMOSA", *IEEE Transaction on Evolutionary Computation* (accepted).
- S. Bandyopadhyay and S. Santra, "A Genetic Approach for Efficient Outlier Detection in Projected Space", Pattern Recognition (accepted).
- S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, ``An Improved Algorithm for Clustering Gene Expression Data", *Bioinformatics*, Oxford University Press, vol. 23, no. 21, pp. 2859-2865, 2007.

Applications

- Clustering
  - S. Bandyopadhyay, U. Maulik and A. Mukhopadhyay, “Multiobjective Genetic Clustering for Pixel Classication in Remote Sensing Imagery", IEEE Trans. Geoscience &Remote Sensing, vol. 45, no. 5, pp. 15061511, 2007.
  - S. Bandyopadhyay and U. Maulik, “Non-parametric Genetic Clustering : Comparison of Validity Indices", *IEEE Trans. on Systems, Man and Cybernetics Part-C*, vol. 31, no. 1, pp. 120-125, 2001.
  - U. Maulik and S. Bandyopadhyay, “Performance Evaluation of Some Clustering Algorithms and Validity Indices", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650-1654, 2002.
- Classification
  - S. Bandyopadhyay, S. K Pal and B. Aruna, “Multi-objective GAs, Quantitative Indices and Pattern Classication", *IEEE Trans. on Systems, Man and Cybernetics - B,* vol. 34, pp. 2088-2099, 2004.

Applications contd...

- Computational Biology
  - S. Bandyopadhyay, A. Bagchi and U. Maulik, “Active Site Driven Ligand Design: An Evolutionary Approach", J. of Bioinformatics and Computational Biology, vol. 3, No. 5, pp. 1053-1070, 2005.
  - S. Bandyopadhyay, “An Efficient Technique for Superfamily Classication of Amino Acid Sequences: Feature Extraction, Fuzzy Clustering and Prototype Selection", *Fuzzy Sets & Systems*, vol. 152, pp. 5-16, 2005]
  - S. S. Ray, S. Bandyopadhyay, and S. K. Pal, “Genetic Operators for Combinatorial Optimization in TSP and Microarray Gene Ordering", *Applied Intelligence* (accepted).
  - S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, “An Improved Algorithm for Clustering Gene Expression Data", *Bioionformatics*, Oxford University Press, vol. 23, no. 21, pp. 2859-2865, 2007.

Questions ???

Thank You …