

# Clustering Genetic Algorithm

Petra Kudová

Department of Theoretical Computer Science  
Institute of Computer Science  
Academy of Sciences of the Czech Republic

ETID 2007

# Outline

Introduction

Clustering Genetic Algorithm

Experimental results

Conclusion



# Motivation

## Goals

- study applicability of GAs to clustering
- design genetic operators suitable for clustering
- application to tasks with unknown number of clusters
- compare to standard techniques

## Clustering

- partitioning of a data set into subsets - **clusters**, so that the data in each subset share some common trait
- often based on some similarity or distance measure
- the notion of similarity is always problem-dependent.
- wide range of algorithms (k-means, SOMs, etc.)



# Clustering

## Definition of cluster

- Basic idea: cluster groups together similar objects
- More formally: clusters are connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by an low density of points

## Applications

- **Marketing** - find groups of customers with similar behaviour
- **Biology** - classify of plants/animals given their features
- **WWW** - document classification, clustering weblog data to discover groups of similar access patterns



# Genetic algorithms

## Genetic algorithms

- stochastic optimization technique
- applicable on a wide range of problems
- work with population of solutions - individuals
- new populations produced by genetic operators

## Genetic operators

- **selection** - the better the solution is the higher probability to be selected for reproduction
- **crossover** - creates new individuals by combining old ones
- **mutation** - random changes

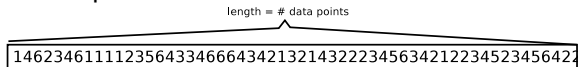


# Clustering Genetic Algorithm (CGA)

## Representation of the individual

### 1. approach (Hruschka, Campelo, Castro)

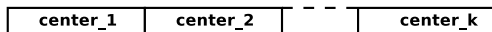
- for each data point store cluster ID



- long individuals (high space requirements)

### 2. approach (Maulik, Bandyopadhyay)

- store centres of the clusters



- need to assign data points to clusters before each fitness evaluation

# Fitness

## Normalization

- partition the data set into clusters using the given individual
- move the centres to the actual gravity centres

## Fitness evaluation

- clustering error:  $fit(l) = -E_{VQ}$

$$E_{VQ} = \sum_{i=1}^K \|\vec{x}_i - \vec{c}_{f(x_i)}\|^2, \quad f(\vec{x}_i) = \arg \min_k \|\vec{x}_i - \vec{c}_k\|^2$$

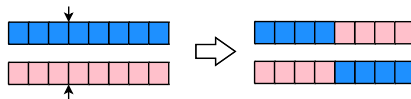
- silhouette function:  $fit(i) = \sum_{i=1}^N s(\vec{x}_i)$

$$s(\vec{x}) = \frac{b(\vec{x}) - a(\vec{x})}{\max\{b(\vec{x}), a(\vec{x})\}}$$

# Crossover

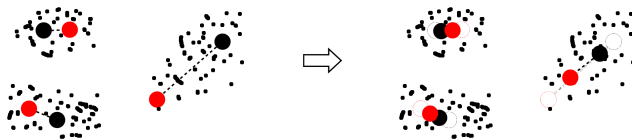
## One-point Crossover

- exchange the whole blocks (i.e. centres)



## Combining Crossover

- match the centres and combine them





# Mutation

## One-point mutation, Biased one-point mutation

- One-point Mutation:

$$\vec{c}_{new} = \vec{x}_i, \text{ where } i \leftarrow \text{random}(1, N)$$

- Bias one-point Mutation:

$$\vec{c}_{new} = \vec{c}_{old} + \vec{\Delta}, \text{ where } \vec{\Delta} \text{ is a random small vector}$$

## K-means mutation

- several steps of k-means clustering

## Cluster addition, Cluster removal

- Cluster Addition* – adds one centre
- Cluster Removal* – removes randomly selected centre

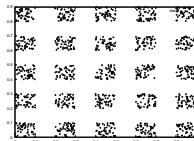
# Experiments

## Goals

- demonstrate the performance of CGA
- compare variants of genetic operators

## Data Sets

- **25 centres**



- **vowels** (UCI machine learning repository)  
11 kinds of vowels, dimension 9  
990 examples
- **mushrooms** (UCI machine learning repository)  
23 kinds of mushrooms, dimension 125  
8124 examples

# Operators Comparison

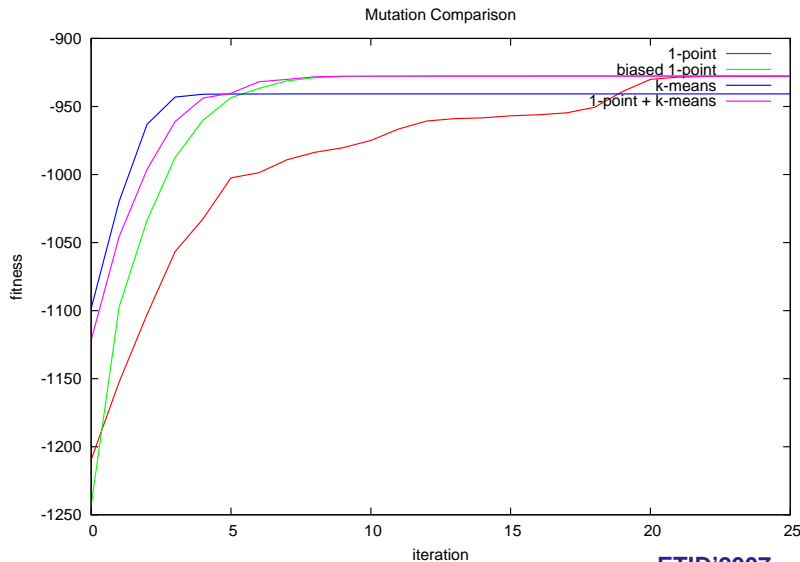
## Mutation

	25clusters	Vowels
1-point	0.20	927.7
Biased 1-point	0.25	927.3
K-means	0.26	940.7
1-point + Biased 1-pt	0.21	927.3
1-point + K-means	0.21	927.6
All	0.22	927.3

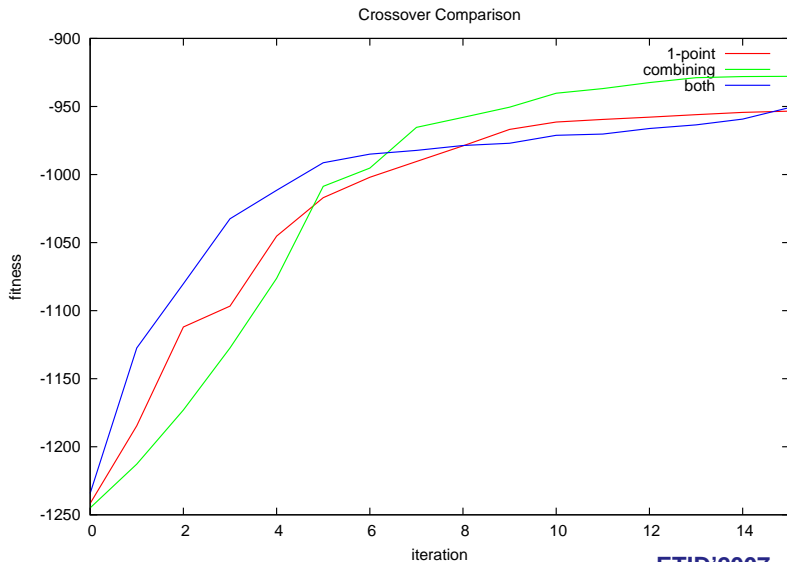
## Crossover

	25clusters	Vowels
1-point	0.201	927.7
Combining	0.222	927.4
Both	0.202	927.4

# Convergence Rate – Mutation



# Convergence Rate – Crossover



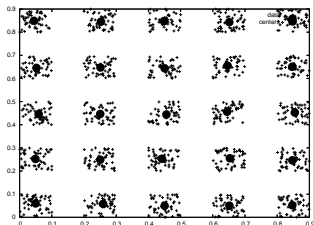
# Comparison to other clustering algorithms

## Mushroom data set

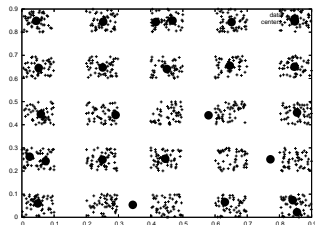
method	accuracy
k-means	95.8%
CLARA	96.8%
<b>CGA</b>	97.3%
HCA	99.2%

## 25 centers

### CGA

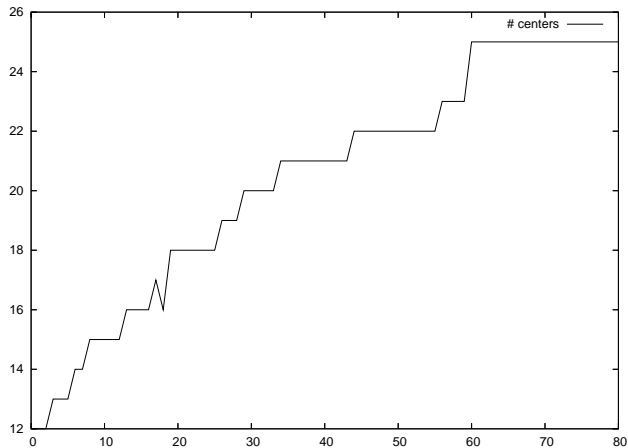


### k-means



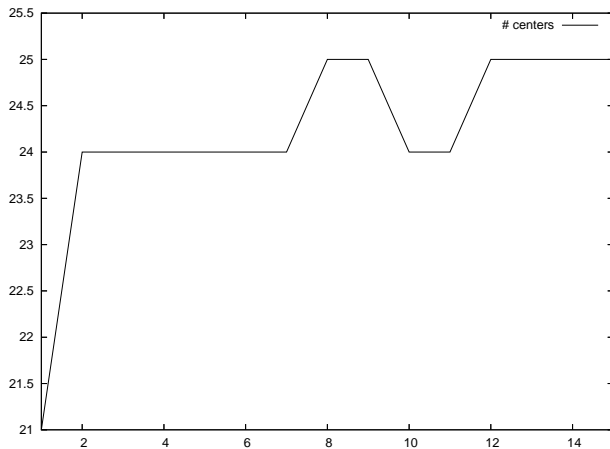
# Estimating the number of clusters

Initial population: 2 to 15 centres



# Estimating the number of clusters

Initial population: 10 to 30 centres





# Conclusion

## Summary

- *Clustering Genetic Algorithm* proposed
- several genetic operators proposed and compared
- CGA compared to available clustering algorithms
- estimating the number of clusters tested

## Future work

- application of CGA to large data sets
- reducing time requirements, lazy evaluations, etc.
- applications



Thank you.  
Any questions?

