Clustering using Genetic Algorithms

Petra Kudová

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

Znalosti 2007

Outline

Introduction

Clustering Genetic Algorithm

Experimental results

Conclusion

Motivation

Clustering

- unsupervised learning (data are unlabelled)
- find structure, clusters
- partition data into subsets that share some common trait

Applications

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

Clustering - problem definition

Goal of 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

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
- Note: The notion of proximity/similarity is always problem-dependent.

Genetic algorithms

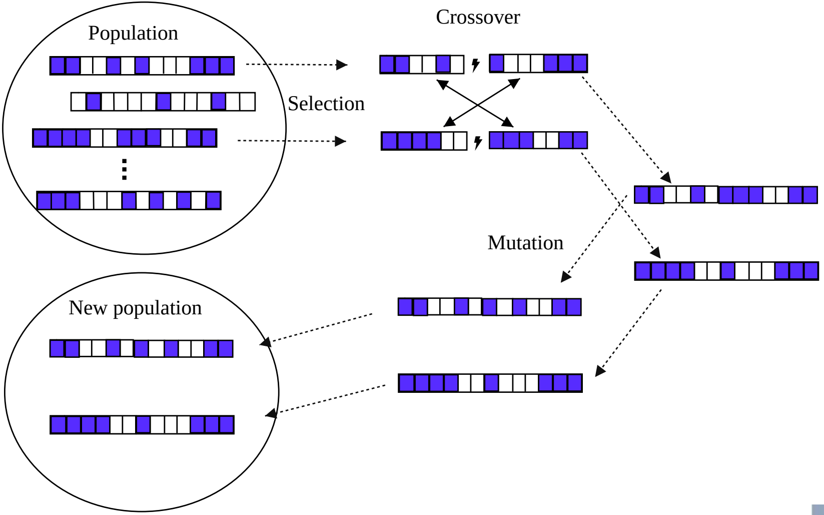
Genetic algorithms

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

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

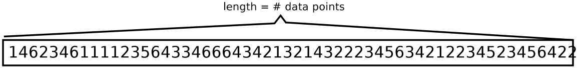
Genetic algorithms



Clustering Genetic Algorithm (CGA)

Representation of the individual

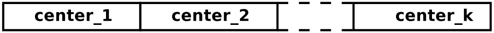
- 1. approach (Hruschka, Campelo, Castro)
  - for each data point store cluster ID


length = # data points

14623461111235643346664342132143222345634212234523456422

  - long individuals (high space requirements), problems in crossover and mutation

- 2. approach (Maulik, Bandyopadhyay)
  - store centres of the clusters

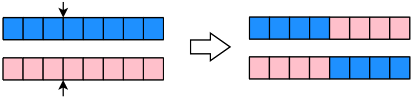

center_1 center_2 center_k

  - need to assign data points to clusters each evaluation

Crossover

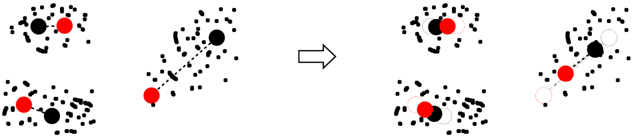
One-point Crossover

- 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:
  $\vec{c}_{new} = \vec{x}_i$, where $i \leftarrow random(1, N)$
- Bias one-point Mutation:
  $\vec{c}_{new} = \vec{c}_{old} + \vec{\Delta}$, where $\vec{\Delta}$ is a random small vector

K-means mutation

- several steps o k-means clustering

Cluster addition, Cluster removal

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

Fitness

Normalization

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

Fitness evaluation

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

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

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

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

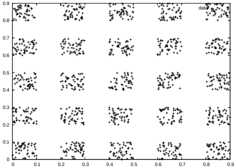
Experiments

Goals

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

Data Sets

- **25 centres**



- **vowels** (UCI machine learning repository)
- **mushrooms** (UCI machine learning repository)

Operators Comparison

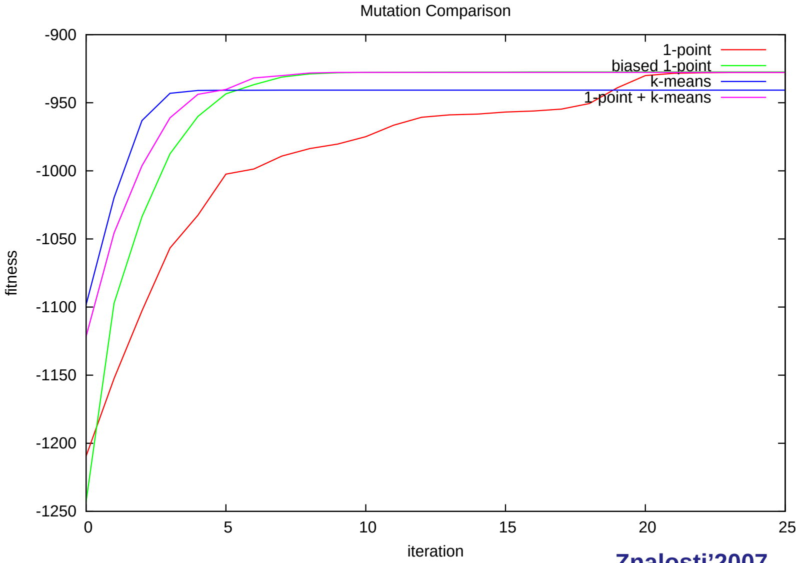
Mutation

	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

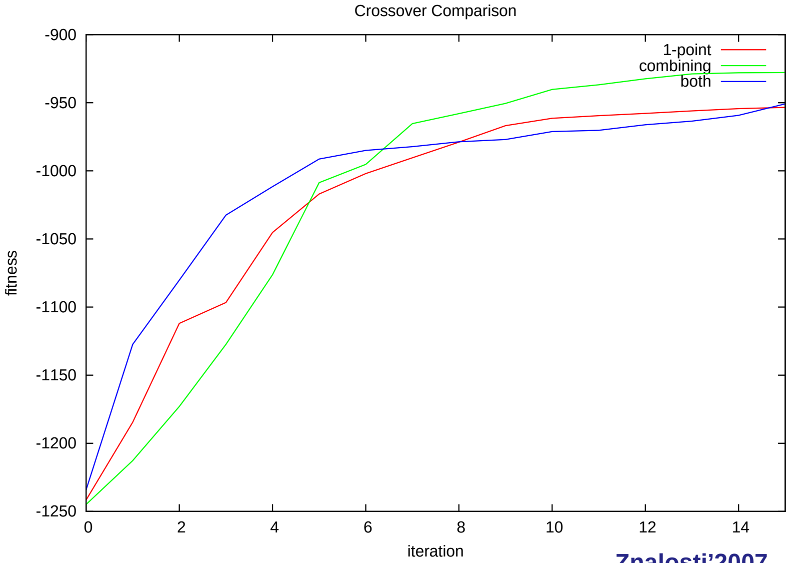
Crossover

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

Convergence Rate – Mutation



Convergence Rate – Crossover



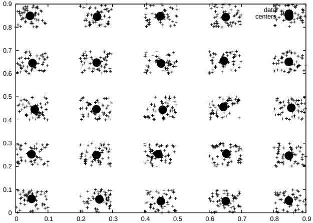
Comparison to other clustering algorithms

Mushroom data set

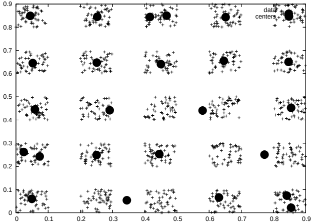
method	accuracy
k-means	95.8%
CLARA	96.8%
CGA	97.3%
HCA	99.2%

25 centers

CGA

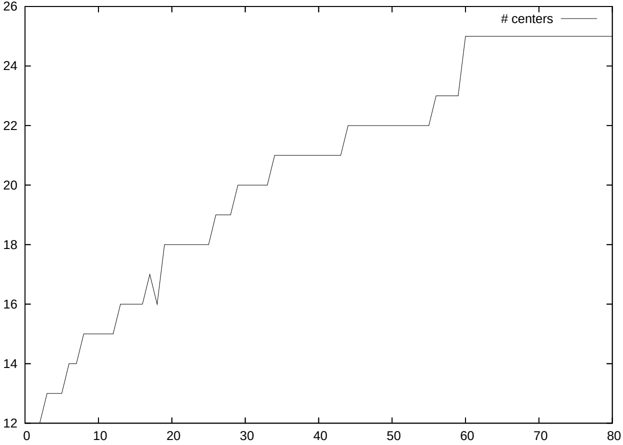


k-means



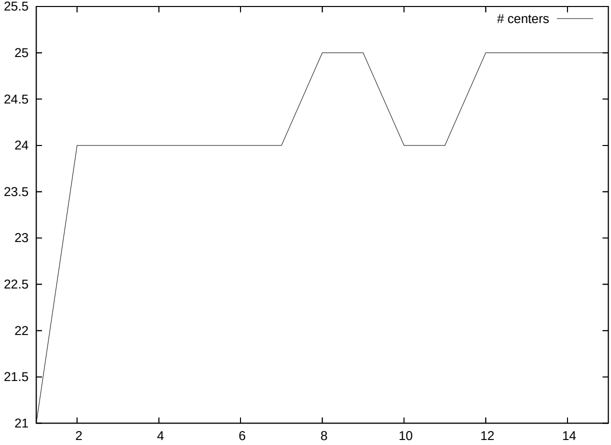
Estimating the number of clusters

Initial population: 2 to 15 centres



Estimating the number of clusters

Initial population: 10 to 30 centres



Conclusion

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

Thank you.
Any questions?