Categorical Data Clustering Using Statistical Methods and Neural Networks

P. Kudová[1], H. Řezanková[2], D. Húsek[1], V. Snášel[3]

[1] Institute of Computer Science
Academy of Sciences of the Czech Republic

[2] University of Economics, Prague, Czech Republic

[3] Technical University of Ostrava, Czech Republic

Outline

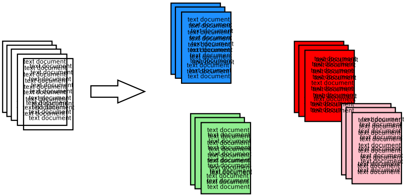
Motivation

Machine learning

- amount of data rapidly increasing
- need for methods for intelligent data processing
- extract relevant information, concise description, structure
- supervised $\times$ unsupervised learning

Clustering

- unsupervised technique
- unlabeled data
- find structure, clusters



Possible applications of clustering

- Marketing - finding groups of customers with similar behavior
- Biology - classification of plants and animals given their features
- Libraries - book ordering
- Insurance - identifying groups of motor insurance policy holders with a high average claim cost, identifying frauds
- Earthquake studies - clustering observed earthquake epicenters to identify dangerous zones
- WWW - document classification, clustering weblog data to discover groups of similar access patterns

Goals of our work

State of the Art

- summarize and study available clustering algorithms
- starting point for our future work

Clustering techniques

- statistical approaches - available in SPSS, S-PLUS, etc.
- neural networks, genetic algorithms - our implementation

Comparison

- compare the available algorithms
- on benchmark problems

Clustering

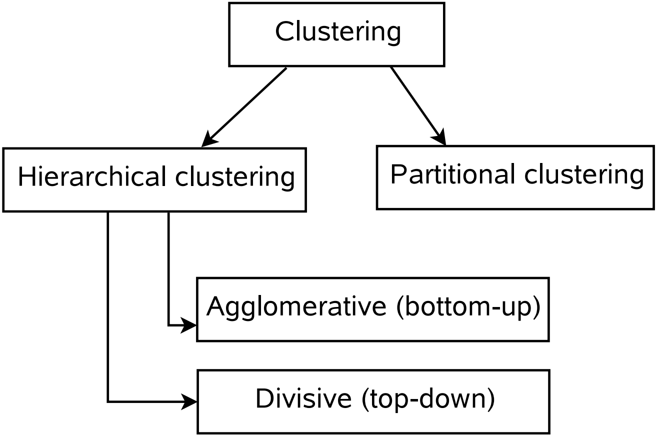
Goal of clustering

- 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.

Overview of clustering methods



Clustering of categorical data I.

Categorical data

- object described by p attributes - $x_1, \ldots, x_p$
- attributes dichotomous or from several classes
- examples: $x_i \in \{yes, no\}$

$$x_i \in \{male, female\}$$

$$x_i \in \{small, medium, big\}$$

Methods for categorical data

- new approaches for categorical data
- new similarity and dissimilarity measures

Clustering of categorical data II.

Problems

- available statistical packages provide similarity measures for binary data
- methods for categorical data rare and often incomplete

Similarity measures

$$s_{ij} = \frac{\sum_{l=1}^{p} g_{ijl}}{p} \qquad g_{ijl} = 1 \iff x_{il} = x_{jl}$$

- Percentual disagreement $(1 - s_{ij})$ (used in STATISTICA)

Clustering of categorical data III.

Similarity measures

- Log-likelihood measure (in Two-step Cluster Analysis in SPSS)
- distance between two clusters $\sim$ decrease in log-likelihood as they are combined into one cluster

$$d_{hh'} = \xi_h + \xi_{h'} - \xi_{\langle h,h' \rangle}; \qquad \xi_g = -n_g \left(\sum_{l=1}^{p} - \sum_{m=1}^{K_l} \frac{n_{glm}}{n_g} log \frac{n_{glm}}{n_g} \right)$$

- CACTUS (CAtegorical ClusTering Using Summaries)
- ROCK (RObust Clustering using linKs)
- k-histograms

Statistical methods

Algorithms overview

- hierarchical cluster analysis (HCA) (SPSS)
- CLARA - Clustering LARge Applications (S-PLUS)
- TSCA - Two-step cluster analysis with log-likelihood measure (SPSS)

Measures used

- Jac Jaccard coefficient - assymetric similarity measure
- CL complete linkage
- ALWG average linkage within groups
- SL single linkage
- ALBG average linkage between groups

Similarity measures

Jaccard coefficient

- assymetric binary attributes, negative are not important

$$s_{ij} = \frac{p}{p+q+r}$$

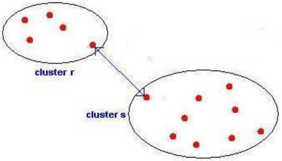
$p \ldots$ # of attributes positive in both objects

$q \ldots$ # of attributes positive only in the first object

$r \ldots$ # of attributes positive only in the second object

Linkage

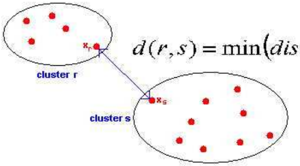
- distance between two clusters



Linkage measures

Single linkage (SL)

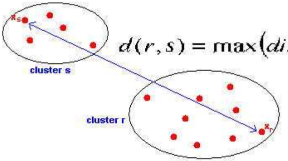
- nearest neighbor



$$d(r,s) = \min(dist(x_{ri}, x_{sj}))$$

Complete linkage (CL)

- furthest neighbor



$$d(r,s) = \max(dist(x_{ri}, x_{sj}))$$

Average linkage(AL)

- average distance

Neural networks and GA

- possible applications of NN and GA on clustering

Neural Networks

- Kohonen self-organizing map (SOM)
- Growing cell structure (GCS)

Evolutionary approaches

- Genetic algorithm (GA)

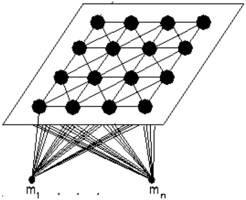
Kohonen self-organizing map (SOM)

Main idea

- represent high-dimensional data in a low-dimensional form without loosing the 'essence' of the data
- organize data on the basis of similarity by putting entities geometrically close to each other

SOM

- grid of neurons placed in feature space
- learning phase - adaptation of grid so that the topology reflect the topology of the data
- mapping phase



Kohonen self-organizing map (SOM) II.

Learning phase

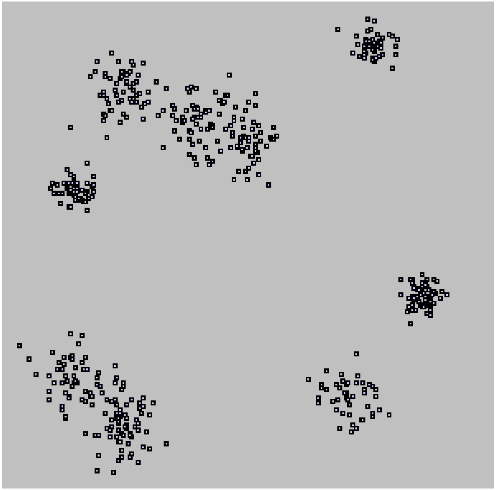
- competition - winner is the nearest neuron
- winner and its neighbors are adapted
- adaptation - move closer to the new point

Mapping of a new object

- competition
- new object is mapped on the winner

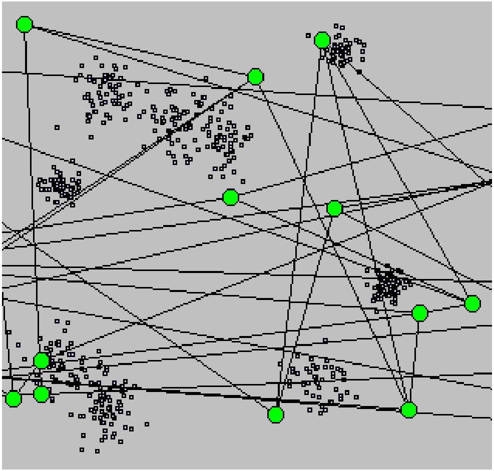
Kohonen self-organizing map (SOM) III.

SOM example



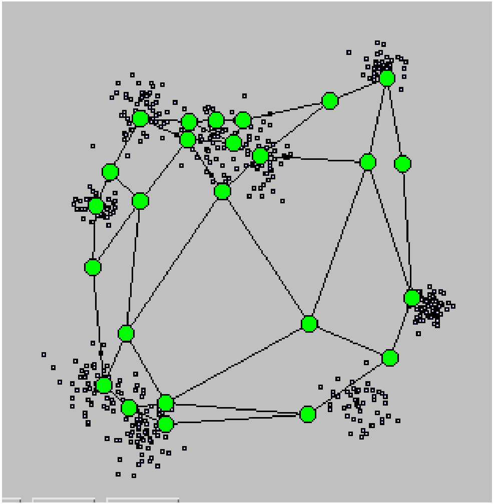
Kohonen self-organizing map (SOM) III.

SOM example



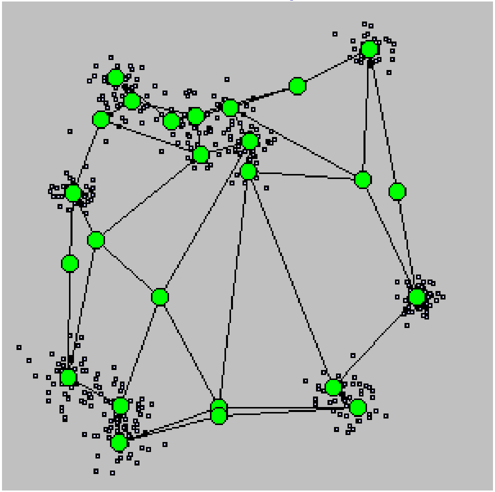
Kohonen self-organizing map (SOM) III.

SOM example



Kohonen self-organizing map (SOM) III.

SOM example



Growing cell structures (GCS)

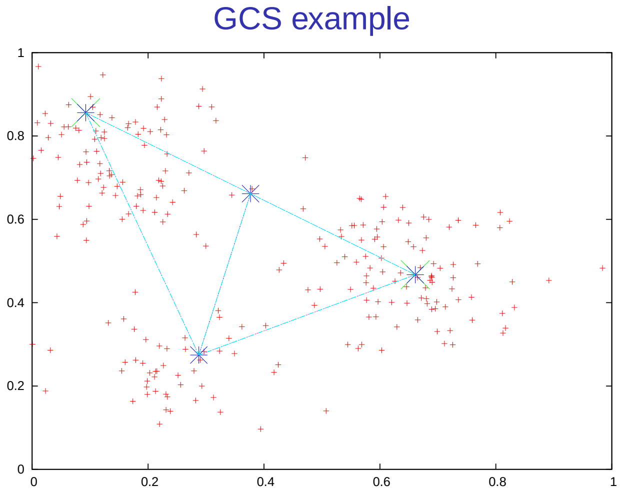
Network topology

- derivative of SOM
- grid - not regular
- network of triangles (or k-dimensional simplexes)

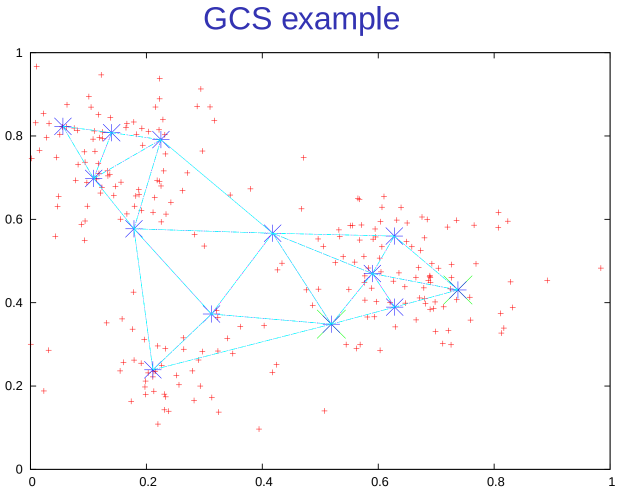
Learning

- learning similar to SOM
- new neurons are added during learning
- superfluous neurons are deleted

Growing cell structures (GCS) II.

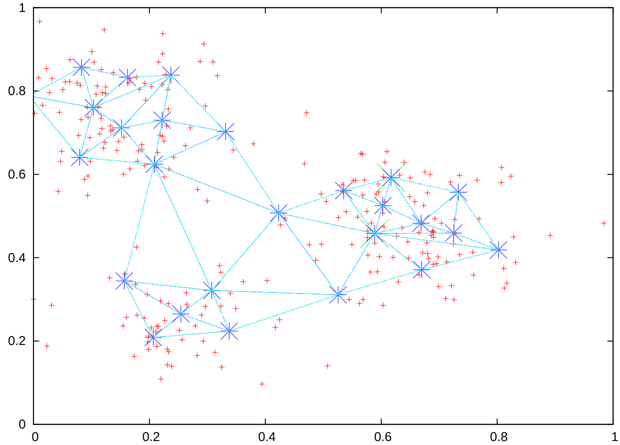


Growing cell structures (GCS) II.

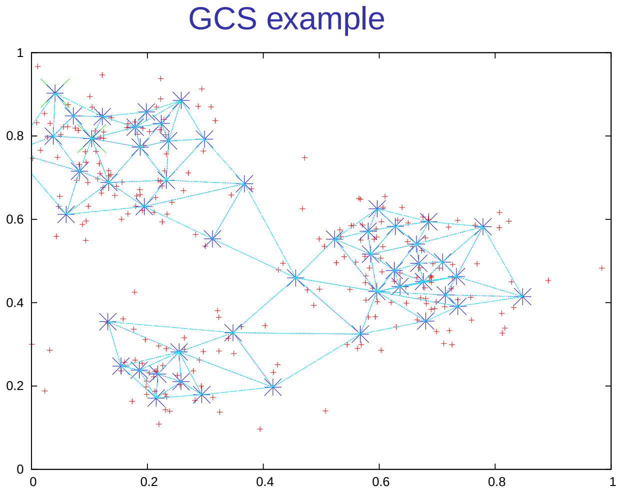


Growing cell structures (GCS) II.

GCS example



Growing cell structures (GCS) II.



Genetic algorithms (GA)

GA

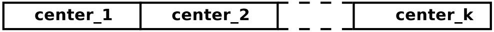
- stochastic optimization technique
- applicable on a wide range of problems
- work with population of solutions - individuals
- new populations produced by operators selection, crossover and mutation

GA 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

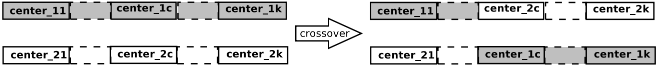
Clustering using GA

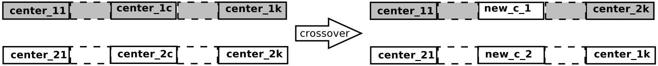
Individual



$$E = \sum_j ||x_j - c_s||^2; \qquad c_s \ldots \text{nearest cluster center}$$

Operators





new_c_1 = A*center_1c+(1-A)*center_2c
new_c_2 = (1-A)*center_1c+A*center_2c

Experimental results

Data set

- Mushroom data set - available from UCI repository
- popular benchmark
- 23 species
- 8124 objects, 22 attributes
- 4208 edible, 3916 poisonous

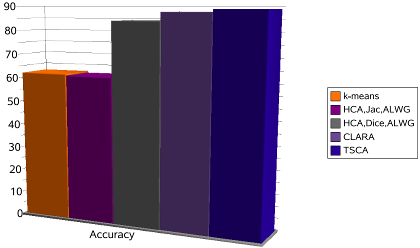
Experiment

- compare different clustering methods
- clustering accuracy

$$r = \frac{\sum_{v=1}^{k} a_v}{n}$$

Statistical methods - 2 clusters

	Edible		Poisonous		Accuracy
Method	Correct	Wrong	Correct	Wrong	
k-means	3836	372	1229	2687	62.3%
HCA, Jac, ALWG	3056	1152	1952	1964	61.6%
HCA, Dice, ALWG	3760	448	3100	816	84.4%
CLARA	4157	51	2988	928	87.9%
TSCA	4208	0	3024	892	89.0%



Number of “pure” clusters

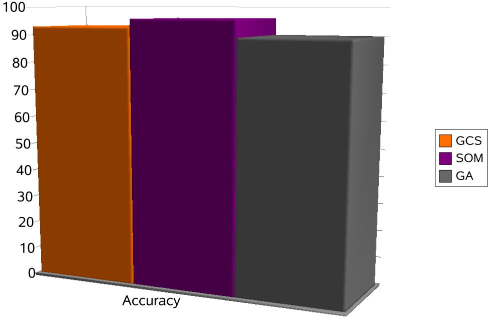
	Total number of clusters							
Methods	2	4	6	12	17	22	23	25
k-means	0	0	0	2	9	16	16	16
HCA, Jac, CL	0	2	2	9	15	20	21	23
HCA, Jac ,ALWG	0	1	2	7	12	18	19	21
HCA, Jac, ALBG	1	2	3	8	13	21	23	25
HCA, Jac, SL	1	3	4	10	14	22	23	25
TSCA – binary	1	3	4	8	14	20	21	24
TSCA – nominal	1	3	4	8	14	20	21	22
CLARA	0	0	0	7	7	13	15	16

Accuracy for different number of clusters

	Total number of clusters						
	4	6	12	17	22	23	25
k-means	78%	80%	92%	94%	95%	95%	98%
HCA, Jac, CL	76%	82%	97%	98%	98%	99%	99%
HCA, Jac, ALWG	88%	88%	95%	98%	99%	99%	99%
HCA, Jac, ALBG	68%	89%	89%	94%	99%	100%	100%
HCA, Jac, SL	68%	89%	89%	91%	100%	100%	100%
CLARA	90%	75%	93%	96%	93%	96%	98%
TSCA – binary	89%	89%	95%	97%	98%	99%	99%
TSCA – nominal	89%	89%	93%	98%	99%	99%	99%
GCS	x	90%	92%	90%	93%	91%	95%

Neural Networks and GA

Method	Accuracy	# clusters
GCS	93%	22
SOM	96%	25
GA	90%	4



Conclusion

Statistical methods and Neural networks

- statistical methods give better accuracy
- GCS, SOM provide topology, not only clustering
- GA good accuracy with 4 clusters , but time consuming

Future work

- focus on hierarchical methods
- clustering using kernel methods
- application, clustering text documents