Evolutionary k-means clustering method with controlled number of detected groups applied in determining the typology of Polish municipalities

Jarosław Stańczak and Jan. W. Owsiński

Systems Research Institute
Polish Academy of Sciences

Introduction

- The widely-known k-means algorithm is a very good tool for obtaining clustering of data.
- It is rather fast and exact method, although its theoretical computational complexity is NP-hard.
- Unfortunately it is difficult to find the proper number of clusters, which is the most important parameter of this method, imposed by its user.
- The presented k-means evolutionary method is a proposed method of solving this drawback.

The k-means algorithm

The basic algorithm of k-means can be described as follows:

1. Choose the number of sought clusters.
2. Generate starting positions of cluster centroids
3. Calculate distances of all clustered objects to all cluster centroids.
4. Assign objects to clusters with the closest centroids.
5. Update cluster centroids as geometric centers of their clusters.
6. If the assignment of objects to clusters in the subsequent two steps does not change, then go to 7, else go to 3.
7. End.

The k-means algorithm

Generally the k-means algorithm minimises the criterion being the sum of distances between each point and its the closest cluster center:

$$C_D(P) = \Sigma_q \Sigma_{i \in Aq} d(x_i, x^q), \qquad (1)$$

where $d(.,.)$ - denotes the squared Euclidean distance; in a more general setting Minkowski distance can also be used; x_i – clustered data items; x^q – centroids of clusters A_q, q=1 ... p^{max}, p^{max} – is the imposed number of detected clusters.
The Minkowski distance is defined as:

$$d(x_i,x_j) = (\Sigma_k (x_{ik} - x_{jk})^h)^{1/h}, \qquad (2)$$

where k denotes the index of coordinates/attributes, describing the processed data x_i, x_j; i, j - data indexes; h is an exponent, h>0. Convergence of k-means is proven, though, only for squared Euclidean distance.

The k-means algorithm

- Unfortunately the criterion (1) cannot be the base of determining the proper („optimal”) number of clusters in processed data set P.
- This is because the formula (1) reaches its global minimum value equal 0 in the case of imposing the number of clusters (p^{max}) equal to the number of processed data |P|.
- In this case each data point becomes a center of its own cluster and its distance $d(x_i,x^q)$ to this center equals 0. This case is of course not very useful, sought values are between 1 and |P|.
- Thus, the criterion and the method of seeking the proper or „optimal” number of clusters should be changed.
- An evolutionary algorithm with a little modified criterion (1) is our proposition, presented in the following slides.

The standard evolutionary algorithm (EA)

The standard evolutionary algorithm works as this is shown below:

1. Random initialization of the population of solutions.
2. Reproduction and modification of solutions using genetic operators.
3. Evaluation of obtained solutions.
4. Selection of individuals for the next generation.
5. If stop condition is not satisfied go to 2, else go to 6.
6. End.

The evolutionary algorithm

specialization

The standard evolutionary algorithm requires usually several modifications to work efficiently:

- invention of proper encoding of solutions;
- development of specialized genetic operators;
- preparing of fitness function (modified problem's criterion);
- adopting a selection method;
- determining the number of iterations to be performed.

The specialized evolutionary algorithm

problem encoding

- number of detected clusters,
- values of centroids of clusters,
- the exponent in Minkowski distance (optional, can be imposed by the user, in the presented case set to 2),
- weights of data attributes (optional, can be imposed by the user, in the presented case all weghts are set to 1),
- value of *r* - described later „zooming” parameter (optional, can be imposed by the user from the interval $\langle 0,1 \rangle$).

The specialized evolutionary algorithm

specialized genetic operators

- mutation that modifies the number of centroids,
- mutation that modifies values of centroids,
- mutation that modifies other optional parameters (not used in the presented case),
- averaging crossover (weighted average parameter values from crossed solutions),
- uniform crossover (exchange of parameters between solutions).

A special method of management of execution of genetic operators that is based on machine learning has been used.

The specialized evolutionary algorithm

selection method

- Usually a typical selection method is used in evolutionary algorithms, the most popular is the tournament selection.
- In our solution a specialized controlled selection method is used, which consists of 2 methods:
  - histogram selection which has a weak selection pressure but increases the population diversity,
  - deterministic roulette selection which has a strong selection pressure but easily unifies the population.
- The controlled selection method tries to maintain the population diversity and preserve the strong selectiom pressure to speed up evolutionary computations.

The maximum number of iterations is is selected experimentally – 10 000 iterations.

The specialized evolutionary algorithm

fitness function

$$C_{Dr}(P) = \Sigma_q \Sigma_{i\in Aq} d_r(x_i, x^q), \tag{3}$$

where: $d_r(x_i,x^q)$ – denotes modified Euclidean/Minkowski distance,
x_i – clustered data, x^q – centroids of clusters A_q, q=1 ... p^{max}.

The modified Euclidean/Minkowski distance is calculated as follows:

If $d(x_i,x^q) \le R$ then $d_r(x_i,x^q)=R$, in the opposite case $d_r(x_i,x^q)=d(x_i,x^q)$,

The R value is calculated based on the properties of the grouped data and the given parameter r, $r \in \langle 0,1 \rangle$, which is meant to control the degree of detail of the clustering:

$$R = (1-r)*0.2*d_{min}(x_i,x_j)+r*0.8*d_{max}(x_i,x_j), \tag{4}$$

where: $d_{min}(x_i,x_j)$ is the minimum value (but bigger than zero) of the Euclidean/Minkowski distance among grouped data, while $d_{max}(x_i,x_j)$ is the maximum value of the Euclidean/Minkowski distance among grouped data.

The evolutionary k-means method

1. Random initialization of the population of solutions.
2. Reproduction and modification of solutions using genetic operators.
3. Evaluation of obtained solutions:
   a) total minimized distance (3) is equal to infinity, the number of steps is equal to 0
   b) take the number and centers of sought clusters from evaluated solution,
   c) calculate distances (meant as in formula (2)) of all clustered objects to all cluster centroids,
   d) assign objects to clusters with the closest centroids,
   e) update cluster centroids as geometric centers of their clusters,
   f) if calculated total distance for new data clustering (3) is less than calculated in previous step and number of steps is less than 5, then go to b).
4. Selection of individuals for the next generation.
5. If stop condition of EA is not satisfied go to 2, else go to 6.
6. End.

Typology of Polish municipalities

No.	Attribute	No.	Attribute
1	Population	12	Index of the average area of an agricultural holding
2	Built-up area	13	Share of registered working people
3	Share of transport areas	14	Number of registered economic activities per 1,000 inhabitants
4	Density of population	15	Average employment rate in operating enterprises
5	Share of agricultural land	16	The share of enterprises from industry and construction
6	Share of built-up land	17	Number of pupils and students per 1,000 inhabitants
7	Share of forest areas	18	Number of pupils and students of secondary schools per 1,000 inhabitants
8	Share of population over 60 years of age	19	Own income of the commune per capita
9	Share of the population below 20 years of age	20	Share in PIT as part of the commune's budget
10	Birth rate in the last 3 years	21	Share of expenses for social purposes in the commune's budget
11	Migration balance in the last 3 years		

Attributes of data describing municipalities.

Typology of Polish municipalities

Functional types	Number of units		Population		Area		Population density
	No.	%	in ‘000	%	‘000 km^2	%	Persons/sq. km
1. Functional urban areas of voivodship (provincial) capitals	33	1.3	9 557	24.8	4.72	1.5	2 025
2. External zones of provincial capitals	266	10.7	4 625	12.0	27.87	8.9	166
3. Functional urban areas of subregional centers	55	2.2	4 446	11.6	3.39	1.1	1 312
4. External zones of subregional centers	201	8.1	2 409	6.3	21.38	6.8	113
5. Multifunctional urban centers	147	5.9	3 938	10.2	10.39	3.3	379
6. Communes with developed transport functions	138	5.6	1 448	3.8	20.06	6.4	72
7. Communes with developed other non-agricultural functions	222	9.0	1 840	4.8	33.75	10.8	55
8. Communes with intensively developed agricultural functions	411	16.6	2 665	6.9	55.59	17.8	48
9. Communes with moderately developed agricultural functions	749	30.2	5 688	14.8	93.83	30.0	61
10. Communes featuring extensive development (with forests or nature protection areas)	257	10.4	1 878	4.9	41.59	13.3	45
Totals for Poland	2 479	100	38 495	100	312.59	100	123

Functional typology of Polish municipalities.

Obtained results of computer simulations

r	$C_{Dr}(P)$	Number of detected clusters (p)
0.00	577.09	60
0.10	630.58	57
0.20	1038.59	46
0.30	1520.13	31
0.40	2004.66	15
0.45	2246.92	10
0.50	2489.17	7
0.60	2973.69	5
0.70	3458.20	3
0.80	3942.75	4
0.90	4427.23	2
1.00	4911.64	1

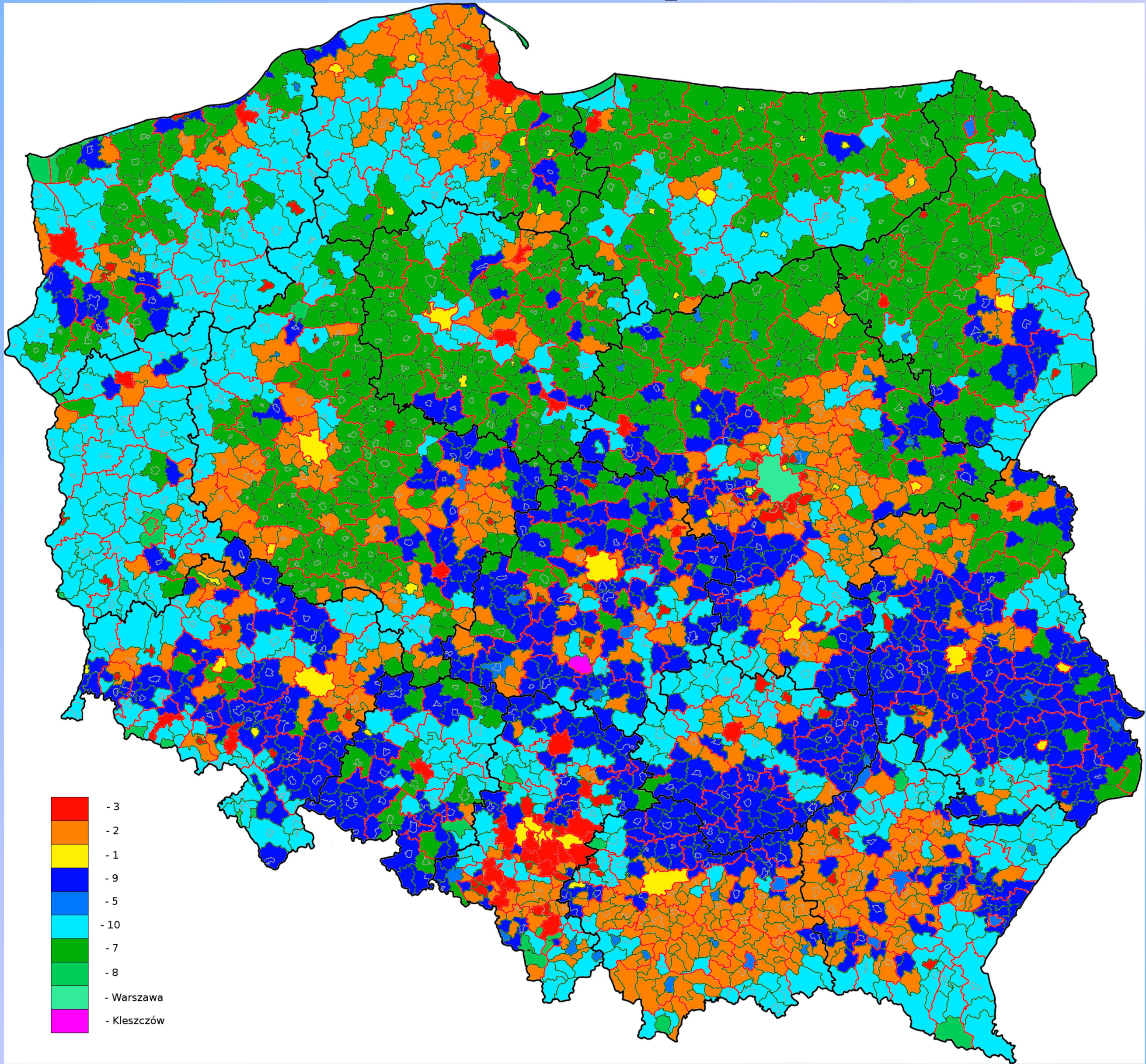
Results showing the dependence of the number of obtained clusters on imposed value of parameter r for data on Polish communes.

Obtained results of computer simulations

Categories: computed→ given↓	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Totals**
1	15	0	**14**	0	1	0	0	2	1	0	33
2	29	**98**	9	54	3	49	19	4	0	0	265
3	**26**	0	14	0	12	0	0	3	0	0	55
4	6	62	0	55	1	52	20	5	0	0	201
5	43	7	15	15	**36**	8	13	5	0	0	142
6	3	23	0	49	2	30	30	0	0	0	137
7	4	30	2	26	2	97	**44**	16	0	1	222
8	0	22	0	141	0	0	0	**333**	0	0	496
9	0	190	0	**258**	8	125	84	0	0	0	665
10	2	46	0	43	1	**141**	27	2	0	0	262
Totals	128	478	54	641	66	502	237	370	1	1	2478

Results for the division into 10 categories (r=0.45) for Polish communes (the values in bold are an attempt to manually assign clusters computed vs. given by experts).

Obtained results of computer simulations



The obtained division of communes into 10 categories (***r***=0.45) using the k-means method for the entire Poland (with the interpretation according to table presented on previous slide).

Conclusions

- The presented algorithm shows interesting properties and thus a great potential for use in pre-grouping of data.
- Our further works should concentrate on including more information about grouped data and more efficient fusion of k-means and EA.
- It can be a first step to find more general method of automatic grouping of data.

Thank you !