A concept of multicriteria stratification: a definition and solution

MIKHAIL ORLOV,

DEPARTMENT OF APPLIED MATHEMATICS AND INFORMATICS HSE

BORIS MIRKIN

INTERNATIONAL LABORATORY OF DECISION CHOICE AND ANALYSIS; DEPARTMENT OF APPLIED MATHEMATICS AND INFORMATICS

What is stratification?

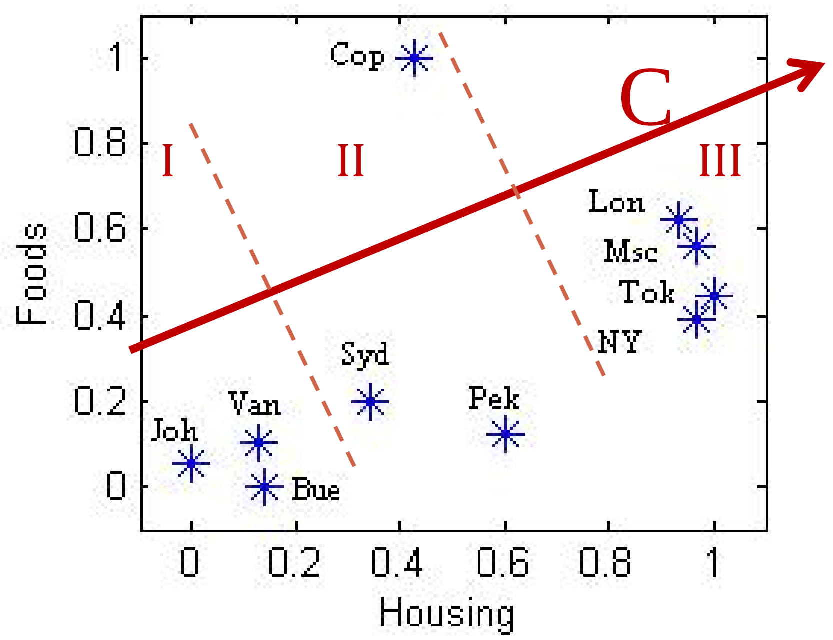
- Geology: “the arrangement of sedimentary rocks in distinct layers (strata)“;
- Sociology: “the hierarchical structures of classes and statuses in any society”.

Stratification example. Food and housing prices

Housing and food prices (2007)
Values are normalized to range to 1.

City	Housing	Foods
Moscow	0.9749	0.7440
London	0.9479	0.7812
Tokyo	1.0000	0.6764
Copenhagen	0.5602	1.0000
New-York	0.9749	0.6446
Peking	0.6924	0.4881
Sydney	0.4967	0.5318
Vancouver	0.3318	0.4775
Johannesburg	0.2322	0.4483
Buenos-Aires	0.3412	0.4178

Aggregate criterion C=aH+bF :
overall expensiveness;
Strata :
I cheap, II medium and III expensive.



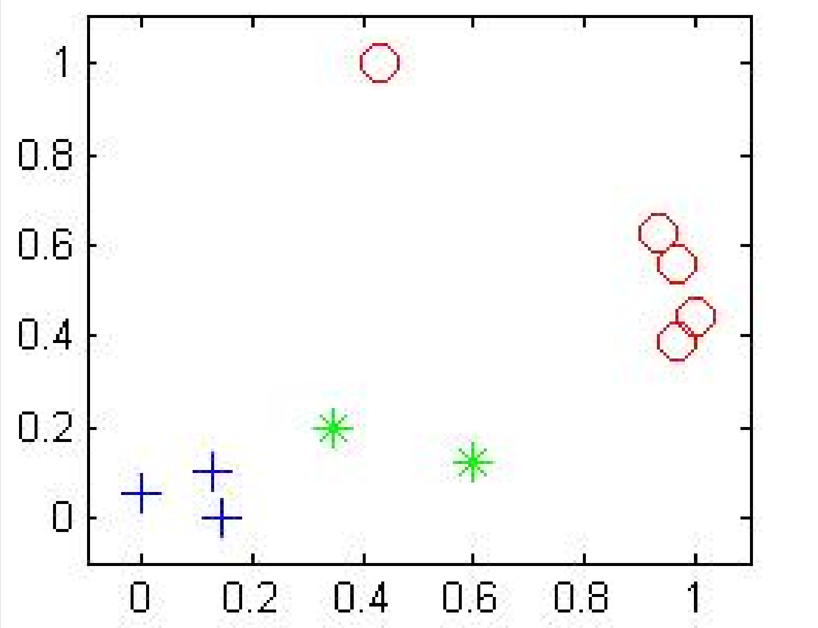
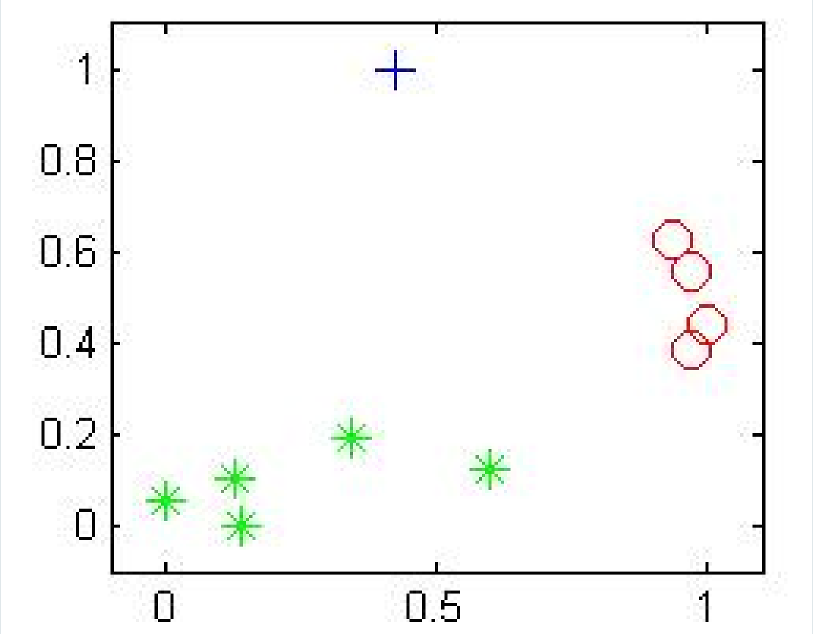
Preliminaries

- N objects are evaluated by M criteria to be maximized;
- Criteria matrix $X = \|x_{ij}\|, i = 1, \ldots, N, j = 1, \ldots, M$;
- Strata are disjoint sets of objects $S = \{S_1, \ldots, S_K\}$;
- Strata are indexed so that the more preferable, the smaller the index.

Problem

- A set of N objects, evaluated by M criteria, should be assigned with **an aggregate criterion W and split into *K* disjoint ordered subsets (strata)** so that W-values in the same group are as close to each other as possible.

Distinction between strata and clusters

Strata	Clusters
	

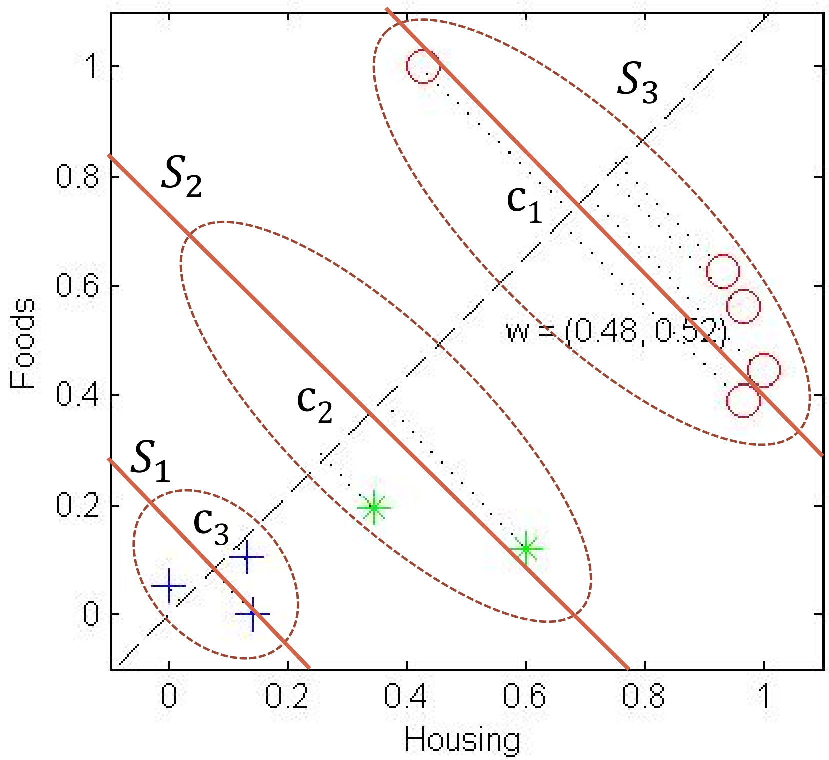
Proposed model for strata

- If object x_i belongs to stratum S_k then:

$$\underbrace{x_{i1}w_1 + x_{i2}w_2 + \cdots + x_{iM}w_M}_{\text{Aggregate criterion value}} = c_k + e_i$$

- w – vector of weights of criteria;
- c_k–center or level of k-th stratum, $c_k \in \{c_1, \dots, c_K\}$;
- e_i - error to be minimized.

Strata in the cities example



Linear stratification criterion

- The problem of stratification:

$$\begin{cases} \sum_{k=1}^{K} \sum_{i \epsilon S_k}^{N} \left(\sum_{j=1}^{M} x_{ij} w_j - c_k \right)^2 \xrightarrow[w,c,S]{} min \\ \sum_{j=1}^{M} w_j = 1, w_j \geq 0 \end{cases}$$

Related work

- Weighted sum of criteria [Sun et al 2009], [Ng 2007; Ramanathan 2006];
- Multicriteria rank aggregation [Aizerman, Aleskerov 1995; Mirkin 1979];
- Multicriteria decision analysis, outranking [DeSmet, Montano, Guzman 2004], [Nemery, DeSmet 2005];

Why do we need stratification at all?

- Expert opinion is often a scale with few grades. E. g. 3-graded: “Good”, “Medium” and “Bad”, or ABC grades;
- Complete order of many items can be inconvenient to work with: choosing a university program according to some rating. What is the point to prefer 500-th item to 501-th out of a thousand?

Computational comparison: Data specification

- **A model for generating synthetic data sets;**
- **Two real datasets;**
- **Two types of criteria normalization:**
  - statistical (scaling to zero mean and unity std.)
  - standard (scaling to the range 0 to 1).

Synthetic data sets

Examples of 3-strata artificial datasets generated by our model.

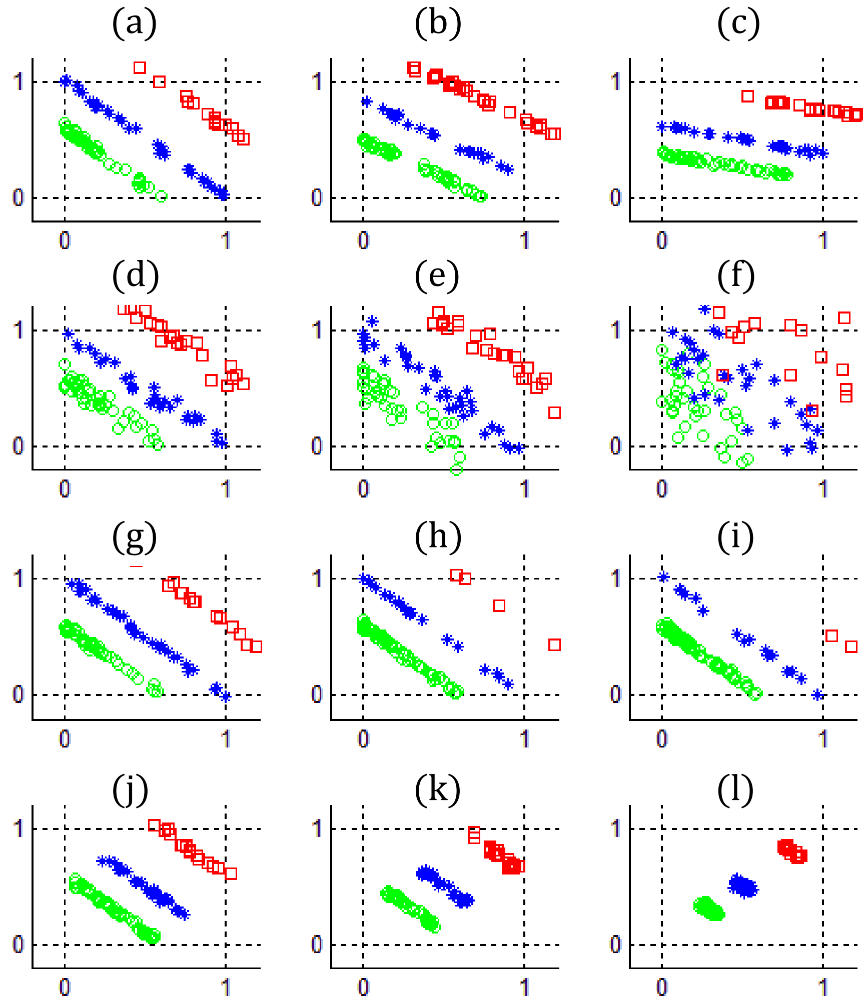
Parameters :

(a),(b),(c) – orientation;

(d),(e),(f) – thickness;

(g),(h),(i) – intensities;

(j),(k),(l) – spread.



Real dataset 1

- Bibliometric indexes for 118 scientific journals in Artificial Intelligence, 2012 [from SCImago Journal & Country Ranking Database]:

  - Index SJR (Scientific Journal Ranking);

  - Hirsch index (number of documents that received at least h citations);

  - Impact-factor.

Real dataset 2

- **Bibliometric indexes of 102 countries at 2012, in Artificial Intelligence:**

  - Total number of documents published in 2012;

  - Number of citable documents published in in 2012;

  - Citations received in 2012 for documents published the same year;

  - Country self-citations in 2012;

  - Citation per document in 2012;

  - Country Hirsch index.

Methods under comparison

- Algorithms for optimization the linear stratification criterion:

  -Evolutionary minimization [Mirkin, Orlov 2013];

  -Quadratic programming [Orlov 2014].
- Rankings partitioned using k-means:

  - Borda count;

  - Linear weight optimization [Ramanathan (2006)];

  - Authority ranking [Sun et. Al 2009].
- Pareto layers merged using agglomerative clustering:

  - Pareto stratification [Mirkin, Orlov 2013].

Evaluation criteria

- On synthetic data. Stratification accuracy:

$$accuracy = \frac{N_{correct}}{N}$$

- On real data. Coherence of obtained stratification with respect to stratifications over single criteria using Kemeny-Snell distance:

$$d_{RS} = \frac{1}{2N(N-1)} \sum_{i,j=1}^{N} |R_{ij} - S_{ij}|$$

$$S_{ij} = \begin{cases} 1, S(x_i) > S(x_j) \\ 0, S(x_i) = S(x_j) \\ -1, S(x_i) < S(x_j) \end{cases}$$

Experimental results on synthetic data

- Accuracy of stratification with respect to the following data generation parameters:
  - data dimensionality,
  - number of objects,
  - strata “intensities”,
  - “spread”,
  - “thickness”.
- In most cases our quadratic programming based algorithm LSQ demonstrated the best accuracy.

Real data set 1 (3 strata)

- In the first stratum:

  1. IEEE Transactions on Pattern Analysis and Machine Intelligence (United States);

  2. International Journal of Computer Vision (Netherland);

  3. Foundations and Trends in Machine Learning (United States);

  4. ACM Transactions on Intelligent Systems and Technology (United States);

  5. IEEE Transactions on Evolutionary Computation (United States);

  6. IEEE Transactions on Fuzzy Systems (United States).

- Criteria weights:

  - Impact Factor: 0.47;

  - Scientific Journal Ranking (SJR): 0.38;

  - Hirsch Index: 0.05.

Real data set 2 (3 strata)

- The first stratum consists of two countries: China, USA.
- The second stratum, 17 countries: Spain, UK, France, Taiwan, Japan, India, Germany, Canada, Italy, South Korea, Australia, Hong-Kong, Netherlands, Singapore, Switzerland, and Israel.
- The other 83 countries form the 3-rd strata.
- Non zero weights:

  - Self-citation: 0.52;

  - Hirsch-index : 0.41;

  - Average citation number: 0.07.

Conclusion

- The problem of multicriteria stratification is formalized as an optimization task to minimize the thickness of strata;
- Two algorithms are proposed;
- A stratified synthetic data generating algorithm is proposed;
- In most synthetic data cases our QP algorithm demonstrated superior performance;
- Application of methods to real data leads to sensible results.

Future work

- Avoiding trivial solutions: If some of criterion is k-valued then optimization task has a trivial minimum. Just assign weight 1 to this feature and get a solution;
- Extensive experimental study of the developed and existing stratification methods on real world data sets;
- Probabilistic formulation of strata model;
- Choosing right number of strata;
- Interpretation of stratification results .

References

Aleskerov F., Pislyakov V., Subochev A. (2013) Rankings of economic journals constructed by the Social Choice Theory methods: Working paper WP7/2013/03. National Research University "Higher School of Economics". – Moscow : Publishing House of the Higher School of Economics, 2013. – 48 p. (in Russian).

Aleskerov F., Khabina E., Schwartz D. (2006) Binary relations, graphs and group decisions. Moscow HSE, 2006. (in Russian)

Belov V., Korichneva J.. (2012) Multicriteria ABC-classification. Quality criteria and canonical algorithms. Business-informatics. 2012. № 1(19). p. 9–16. (in Russian)

Berzh K. (1962) Graph theory and application. Moscow, 1962. (in Russian)

Mirkin B. (1974) The problem of group choice. Publishing house Nauka. Moscow, 1974. (in Russian)

Mirkin B., Orlov M. (2013) Methods for multicriteria stratification and experimental comparisons: Working paper WP7/2013/06/ National Research University "Higher School of Economics". – Moscow : Publishing House of the Higher School of Economics, 2013. – 32 p. (in Russian)

De Smet Y., Montano Guzman L. (2004) Towards multicriteria clustering: an extension of the k-means algorithm // European Journal of Operational Research. 158. pp. 390-398

DeSmet Y., Gilbart F. (2001) A class definition method for country risk problems. Technical report IS-MG. 2001

Fogel D. B. (1995). Evolutionary Computation. Toward a New Philosophy of Machine Intelligence // IEEE Press. NJ. 1995

Gill P.E., Murray W., Saunders M.A., and Wright M.H. Procedures for Optimization Problems with a Mixture of Bounds and General Linear Constraints // ACM Trans. Math. Software, 1984.

Orlov M. An algorithm for multicriteria stratification (in progress).

Gonzalez Pereira B., Guerrero Bote V., Moya Anegon F. (2010) A new approach to the metric of journals scientific prestige: The SJR indicator // Journal of Informetrics. pp. 379–391

Hirsch Jorge E., (2005). An index to quantify an individual's scientific research output. arXiv.

Kemeny, J. , *Snell, L.* (*1962).* Mathematical Models in the Social Sciences, Ginn, Boston, 145 p.

References

Kennedy J., Eberhart R. C. (2001). Swarm Intelligence. Morgan Kaufmann Publishers. San Francisco. Calif. USA.

MacQueen J. (1967) Some methods for classification and analysis of multivariate observations // Le Cam, J. Neyman (Eds.) 5th Berkeley Symp. Math Statist. Prob. 1. pp. 281–297.

Mirkin B. G. (2012) Clustering: A Data Recovery Approach. CRC Press.

Ng W.L. (2007) A simple classifier for multiple criteria ABC analysis // European Journal of Operational Research. 177. pp. 344–353.

Page L., Brin S., Motwani R., Winograd T. (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

Ramanathan R. (2006) Inventory classification with multiple criteria using weighted linear optimization // Computers and Operations Research. 33. pp. 695-700.

Saaty T.L. (1980) The analytic hierarchy process. McGraw-Hill: New York; 1980.

SCImago Journal & Country Ranking. (2007). SJR — SCImago Journal & Country Rank. (Retrieved January 14, 2014, from http://www.scimagojr.com.)

SCImago Lab. http://www.scimagolab.com/ (retrieved January 14, 2014).

Scopus. http://www.elsevier.com/online-tools/scopus (retrieved January 22, 2014).

Siebelt M., Siebelt T., Pilot P., Bloem R. M., Bhandari M. and Poolman R. W. (2010) Citation analysis of orthopaedic literature; 18 major orthopedic journals compared for Impact Factor and SCImago // BMC Musculoskeletal Disorders.

Spreckelsen C., Deserno T. M. and Spitzer K. (2011) Visibility of medical informatics regarding bibliometric indices and databases // BMC Medical Informatics and Decision Making.

Sun Y., Han J., Zhao P., Yin Z., Cheng H., Wu T. (2009) RankClus: integrating clustering with ranking for heterogeneous information network analysis // Proc. EDBT. 2009. pp. 565-576.

Garfield E. (1994). The Thomson Reuters Impact Factor. Thomson Reuters.

Appendix 1. Proposed algorithm for optimization of the stratification criterion

- Input:
  - Items x_i, $i=1..N$;
  - Number of strata K;
  - Iteration number T;
- Output:
  - Weights w;
  - Strata centers c;
  - Partition S.
- Algorithm linstrat-q:

1. Initialize weights and centers;
2. Given weights and centers find optimal partition:

$$x_i \in S_k, k = argmin_k \left(\sum_{j=1}^{M} x_{ij} w_j - c_k \right)^2, k = 1 \dots K, i = 1..N$$

3. Given weights and partition find optimal centers:

$$c_k = \frac{1}{|S_k|} \sum_{x_i \epsilon S_k}^{N} \sum_{j=1}^{M} x_{ij} w_j$$

4. Given centers and partition find optimal weight from the solution of optimization problem (2).
5. Repeat from 2 until T steps is done.

Appendix 2. Synthetic data generator

- Input:
  - Number of objects N, dimensionality M and number of strata K;
  - Strata centers c;
  - Weights of criteria w;
  - Thickness of strata σ;
  - Intensities of strata θ;
  - Spread of strata φ.

- Output:
- N objects along with
  - Criteria values;
  - Strata indices.

- Algorithm for generating objects stratified:
  1. Sample the stratum index for current object from the multinomial distribution $k \sim M(\theta_1, \theta_2, \ldots, \theta_K)$
  2. Sample value of the aggregate criterion from the Gaussian distribution $r \sim N(c_k, \sigma)$
  3. Generate values of M-1 criteria from the uniform distribution $x_j \sim U(c_k(1-\varphi), c_k(1+\varphi)/w_j), j=1\ldots M-1.$
  4. Compute the last criterion from the stratum hyper plane equation $x_M = (r - w_1x_1 + w_2x_2 + \ldots + w_{M-1}x_{M-1})/w_M.$
  5. Repeat from 1 until N objects are generated.