HEURISTIC OPTIMIZATION TECHNIQUES FOR SOLVING THE OPTIMAL PARTITIONING PROBLEM

Nikola Kaspříková

Abstract

Many data analysis tasks involve solving an optimization problem where the objective function has such properties that the problem can not be conveniently solved using the conventional analytical optimization methods. The problem of a search for optimal grouping of data is addressed in this paper, where the average silhouette width is used as the objective function. An application of two general purpose trajectory based heuristic optimization methods for the analysis of a real world dataset is shown and the behaviour of the algorithms in response to the values of the control parameters is assessed. The behaviour of the local search method, which is a special case of the more general threshold accepting algorithm, is determined by the function which is used for obtaining the next (neighbour) solution candidate and the performance of the threshold accepting algorithm is further influenced by the threshold values. Both these simple trajectory based methods can be easily implemented and the results of the experiments suggest that the algorithms perform well.

Key words: heuristic optimization, local search, threshold accepting

JEL Code: C 61, C 63

Introduction

Many economic optimization problems are difficult to solve analytically (as there may be multiple local optima or there are other undesirable properties) and one has to resort to heuristic techniques, which may reach some suitable solution, even though not necessarily always the optimal one, within reasonable time. Heuristic optimization methods may include trajectory-based methods or the population-based methods (see e.g. Gilli et al. (2011)). Trajectory-based methods work with just one solution, which may be modified in every iteration. Trajectory-based methods include the simulated annealing method, local search or threshold accepting. The population-based methods are methods which work with the population of solutions in every iteration, which is often called a generation. Evolutionary algorithms include the particle swarm optimization or the differential evolution method which was introduced by Storn and Price (1997).

Such computationally intensive methods have recently become popular thanks to the rise of the available computation power. In some cases, the values of the control parameters of these general purpose algorithms have to be tuned for the particular problem at hand (as there does not seem to be the rule for setting the optimal values of these parameters which would be suitable for all the problems). Regarding the differential evolution method, the control parameters of the algorithm include the size of the population, the probability of crossover and the step size. The local search and the threshold accepting methods as the trajectory-based methods require a specification of the way how the next solution to be considered should be obtained. The threshold accepting method further requires the threshold values.

The problem of finding optimal partitioning of multivariate data with respect to particular objective function with the use of selected heuristic algorithms – the local search method and the threshold accepting method - is addressed in this paper. The average silhouette width is used as the objective function and the results of the application of the two simple general purpose heuristic methods are evaluated.

The organization of the paper is as follows: after a brief introduction to the field of unsupervised classification, the principles of the selected heuristic optimization methods, the local search method and the threshold accepting method, are recalled and the description of the data set used in the analysis is provided in the Material and Methods section. Then the results of the analysis are reported.

1 Material and Methods

The dataset used in the analysis has 110 cases and 5 variables and it refers to the performance of undergraduate university students in particular items in the test in mathematics. More details about the data are described in (Kaspříková, 2012), results of an elementary descriptive statistical analysis of examination tests and a prediction model for total score is given in (Kaspříková, 2011) and the analysis is further extended using the IRT models framework in (Kaspříková, 2012b). The data in this paper is analyzed with the aim to learn if there is some natural grouping of students regarding the performance in the test. The variables have comparable scales and no data transformations of the original values were applied for the analysis.

1.1 Clustering

The cluster analysis, or the unsupervised classification task, usually aims at the discovery of the best (whatever it is supposed to mean exactly) groupings of the cases included in the analysis in such a way, that the cases in the same cluster are rather similar (and again the question may be what the (dis)similarity is supposed to mean), whereas the cases from the different clusters are not.

There exist many algorithms for solving this problem; one commonly used clustering of such methods is given by Venables and Ripley (2002). The partitioning (i. e. not hierarchical) methods are considered in this analysis and the alternative methods for the algorithms applied include the methods such as the k-means or the partitioning around medoids.

The results of the clustering may be influenced by the choice of the distance measure. The usual 2 norm is used in this analysis.

1.2 The Objective Function

There exist several reasonable objective functions for evaluation of the quality of the clustering model when the partitioning clustering methods are used. One of the possibilities is to choose the average silhouette width. For the definition of the average silhouette width see the original sources, such as (Rousseeuw et al., 1996).

1.3 The Threshold Accepting and the Local Search Methods

The threshold accepting method has quite broad field of applications, which include e.g. the search for optimal pooling of the deals for the purpose of the credit risk management within the Basel regulatory rules – see (Lyra et al., 2010). This optimization heuristic method has been introduced in (Dueck and Schauer, 1990) and it operates using the following general scheme for minimization of the objective function OF:

1. the initial solution x0 is generated and the current solution xc is assigned the value x0
2. repeat until the limit of the number of iterations is reached:
    1. generate new (neighbour) candidate solution xn
    2. if OF(xc) + t > OF(xn) set xc <-- xn, otherwise keep the current xc
3. return the best xc overall

The control parameters of the threshold accepting algorithm include the (non-negative) threshold value t (which may depend on the iteration number).

The function for obtaining the neighbour (i. e. next candidate) solution has to be chosen for the particular task being solved and it has to be supplied by the user. In this analysis a simple function was used. The solution was represented by four cases of the dataset (initial solution was selected at random) and the remaining cases were then assigned to the most similar case when building the groups. To obtain the neighbour solution, first the component (1st, 2nd, 3rd or 4.) of the current solution to be changed was chosen at random and then the selected component was replaced by another case, chosen at random.

The local search method may be considered to be a special case of the threshold accepting method with the threshold value set to 0.

We use the implementation of the threshold accepting and the local search methods available in NMOF package (Gilli et al. 2011) in the R computing environment (R Core Team, 2013). Since the objective function is minimized in this implementation (as is the usual approach), the average silhouette width (available in the cluster package in R) was always multiplied by -1.

Results

The optimal clustering with up to 4 clusters has been searched for. The solutions with just a single cluster have been penalized through setting the objective function value to 0 for this case. The use of the heuristic methods allows an automatic choice of the most suitable number of groups (2, 3 or 4).

Both the application of the threshold accepting method and the local search method resulted in a two clusters solutions. The values of the control parameters for the algorithms did not have any strong impact on the results. The average silhouette width for the final solution obtained by the local search with 1000 iterations is 0.38055 (see Figure 1 which shows how the objective function value changes with the iteration number).

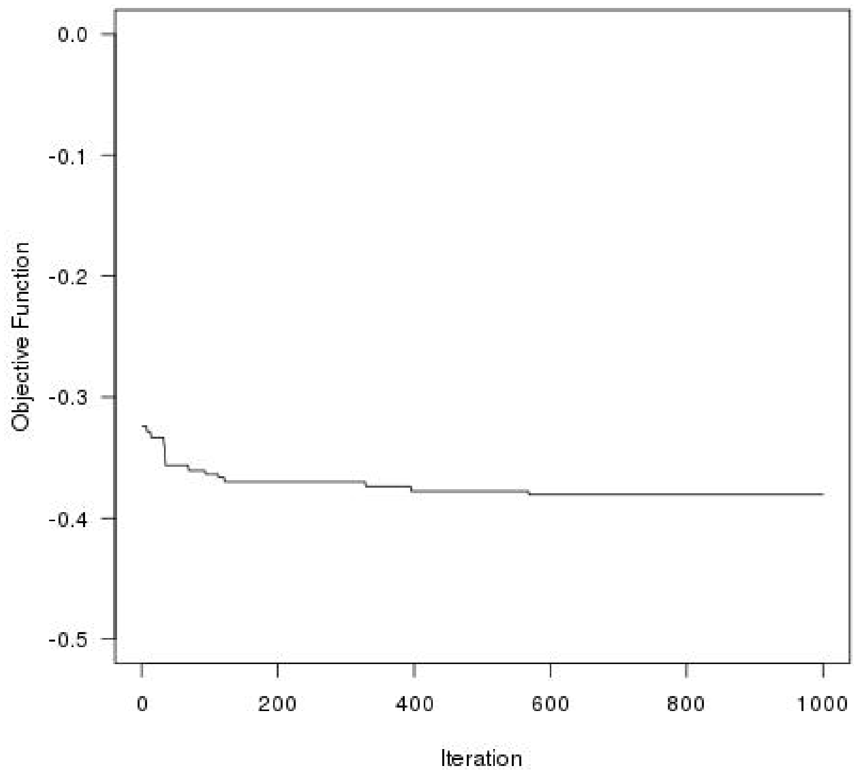
For the threshold accepting with the following parameters:

- the threshold values sequence: *ts* = 0.05, 0.05, 0.025,0.025,0.01,0.01,0.01,0.01,0.01,0.01
- number of iterations per threshold value: 100 , i. e. with 1000 iterations in total,

the average silhouette width is 0.38496 (see Figure 2). So the threshold accepting method did a better job. Nevertheless, the results may be different if other threshold sequence is chosen – for example using the threshold value sequence 0.05,0.05,0.04,0.03,0.02,0.02,0.01,0.01, 0.01,0.01 resulted in the average silhouette width value of just 0.37481.

On the other hand, increasing the number of iterations per threshold value from the original value 100 to 1000, keeping the original threshold values sequence *ts*, allows reaching a better solution – the average silhouette width of such solution is 0.39236 (see Figure 3). But the run time of the computation is significantly higher.

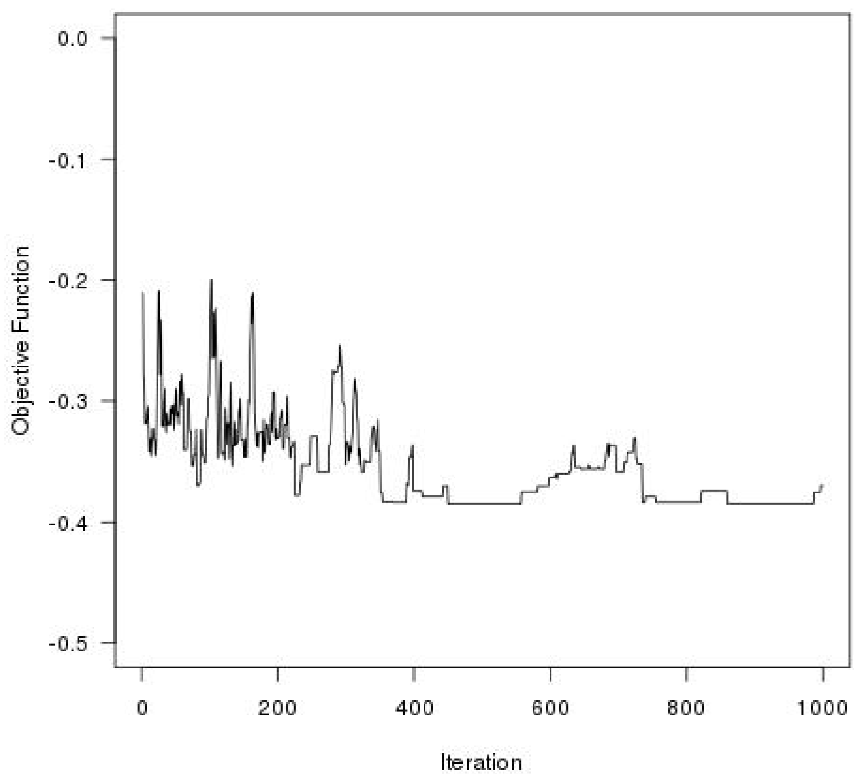
Fig. 1: Local search



Source: own work

Let's consider the partitioning obtained with the threshold accepting algorithm with the threshold values sequence *ts* and 100 iterations per the threshold value as the final partitioning. The final partitioning gives the average silhouette width 0.38496, which is superior to the solution obtained with the standard implementation of the k-means method (in the stats package in R), which gives the average silhouette width for a two clusters solution 0.3790; 0.3371 for 3 clusters solution and 0.2982 for 4 clusters solution. Similarly for the partitioning around medoids method (in the cluster package in R), which gives 0.3079, 0.3388 and 0.2936 respectively. But note that the standard implementations of these algorithms aim at optimizing a different objective.

Fig. 2: Threshold accepting

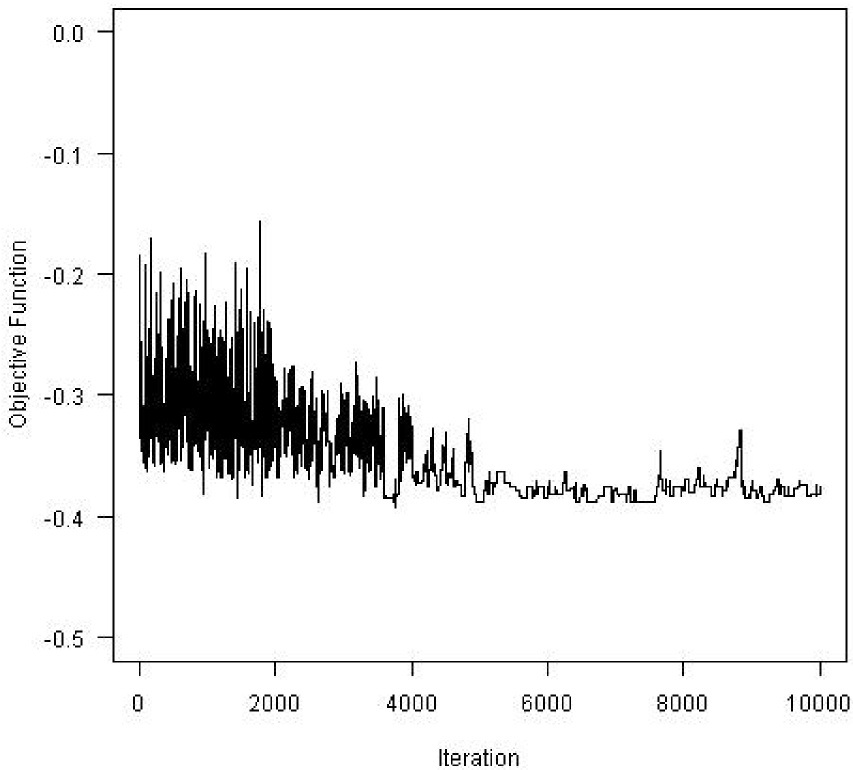


Source: own work

Conclusion

It was shown that both the threshold accepting method and the local search method are suitable for solving the optimal partitioning problem. These methods allow supplying easily the objective function preferred by the user and reach the solution quite quickly. The procedures can automatically determine the best number of clusters, which is also an advantage. The quality of the resulting classification was better when using the threshold accepting method with a good choice of values of the control parameters than when using the local search method.

Fig. 3: Threshold accepting with 1000 iterations per threshold value



Source: own work

References

Dueck, G., & Scheuer, T. (1990). Threshold accepting. a general purpose optimization algorithm superior to simulated annealing. *Journal of Computational Physics, 1 (90)*.

Gilli, M., Maringer, D., & Schumann, E. (2011). Numerical Methods and Optimization in Finance. Academic Press.

Kaspříková, N. (2011). Multivariate Analysis of Examination Papers, *Efficiency and Responsibility in Education*, Proceedings of the 8th International Conference, Czech University of Life Sciences in Prague.

Kaspříková, N. (2012). Data Analysis of Students' Performance. *Efficiency and Responsibility in Education,* Czech University of Life Sciences in Prague.

Kaspříková, N. (2012b). Statistical evaluation of examination tests in mathematics for economists, *Journal of Efficiency and Responsibility in Education and Science, 5 (4).*

M. Lyra, J. Paha, S. Paterlini, & P. Winker (2010). Optimization heuristics for determining internal rating grading scales, *Computational Statistics and Data Analysis, 54*.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousseeuw, P., Struyf, A., & Hubert, M. (1996). Clustering in an object-oriented environment. *Journal of Statistical Software, 1 (4)*.

Storn, R., & Price, K. (1997). Differential Evolution A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, *Journal of Global Optimization 11.*

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Springer.

Contact

Nikola Kaspříková

Department of Mathematics, University of Economics in Prague

Nám.W.Churchilla 4 , 130 67 Praha 3

nb33@tulipany.cz