USING METAHEURISTIC FOR DATA ANALYSIS PROBLEM SOLVING

Nikola Kaspříková

Abstract

Data analysis problems often include some sort of an optimization problem. In many cases, such problems could be efficiently solved using some suitable heuristic optimization technique. The problem of finding the optimal weights for particular variables for distance calculation using data on pairwise distances between subjects is addressed.

The notion of a distance measure is essential in many data analysis tasks. A distance is heavily used in cluster analysis, among others. There are many ways how to define the similarity (or distance respectively). One of the options for a distance definition is based on the Gower's definition of similarity. The weights for particular variables are needed for its calculation.

However, in some situations in practice it may not be possible to get the weights directly.

Results of an application of two simple general purpose heuristic methods are reported. An application of a simple trajectory-based method and a biologically-inspired, population-based metaheuristic method is shown and the performance of the algorithms is discussed.

Key words: distances, metaheuristic, 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 (Kennedy and Eberhart, 1995) or the differential evolution method which was introduced by Storn and Price (1997). Recent development in the field include many new biologically-inspired algorithms, such as the krill herd algorithm (Gandomi and Alavi, 2012).

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 notion of a distance measure is essential in many data analysis tasks. A distance is heavily used in cluster analysis, among others. There are many ways how to define the similarity (or distance respectively). One of the options for a distance definition is based on the Gower's definition of similarity (Gower, 1971). The weights for particular variables are needed for its calculation. However, in many situations in business practice it may not be possible to get the weights directly. But in some cases there may be other data sources available and it is possible to merge information from all the sources available to generate new knowledge which can be used to enhance the business.

The problem of finding the optimal weights for particular variables for distance calculation with respect to particular objective function is addressed in this paper. We seek the solution with the use of selected heuristic algorithms – the local search method, the threshold accepting method and the particle swarm optimization - and using the data on pairwise distances between subjects as a source. The Spearman rank correlation coefficient between the source distances and the distances in the resulting solution is used as the objective function and the results of the application of these general purpose heuristic methods are evaluated.

The structure of this paper is as follows: the distance measure used and the objective function are introduced first and then the principles of the selected heuristic optimization methods, the local search method and the threshold accepting method, are recalled in the Material and Methods section. Then the results of the analysis are reported.

1 Material and Methods

1.1 Data Description

The dataset used in the experiment has 1000 cases and 12 variables. The Gower's distance (for information on distances see e.g. (Gower, 1971), (Venables and Ripley, 2002) or (Rousseeuw et al., 1996) is calculated for each (unordered) pair of cases using a particular chosen weights vector. That means that the solution is going to be represented by a weights vector with 12 elements. In practice, it is sufficient to work with components which have values in interval [0, 1].

1.2 Objective Function

There exist several reasonable objective functions for evaluation of the quality of the solution and the choice of an appropriate function depends heavily on the application problem solved and the ultimate purpose of the analysis. In this experiment, we use the Spearman rank correlation coefficient between the distances calculated using the original variables weight vector (which is known in our experiment, but is often unknown in practice) and the distances calculated using the resulting solution.

1.3 Optimization 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 (denoted 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:
2.1 generate new (neighbour) candidate solution xn
2.2 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 and in this case it means that a sequence of values is required as a configuration parameter of the method).

When working with the local search (or threshold accepting) method, the function for obtaining the neighbour (i. e. the 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 a vector with 12 components (the initial solution used for the computations was selected at random). To obtain the neighbour solution, the values of the current solution are slightly changed by a vector 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. The solution can not deteriorate when using the local search method, whereas when using the general threshold accepting method, the solution can deteriorate (i.e. the objective function value can be worse than at the previous step) to escape a local optimum.

The particle swarm optimization method is a population-based method (several trajectories are considered at a time), for details see the original source (Kennedy and Eberhart, 1995).

We use the implementation of the threshold accepting method, the local search method and the particle swarm optimization method available in the NMOF package (Gilli et al. 2011) in the R computing environment (R Core Team, 2014). Since the objective function is minimized in this implementation (as is the usual approach), the correlation coefficient was always multiplied by -1.

Results

The optimal 12-components weight vector for distance calculation has been searched for in this experiment. The quality of the solutions has been assessed using the Spearman rank correlation coefficient (the coefficient was multiplied by -1 to get the objective function for convenience). The application of any of the selected methods (local search, threshold accepting, particle swarm optimization) resulted in a solution of a high quality.

The local search method with 100 iterations has been applied ten times in the experiment, with various values of the initial solution vector to provide some idea about the impact of the initial control parameters on the quality of the resulting solution. The values of the resulting correlation coefficients have been at least 0.99 for any of the ten runs of the local search algorithm. So the values of the initial solution do not seem to have any strong impact on the results in this case. See Figure 1 which shows how the objective function value changes with the iteration number for a sample run of the local search algorithm.

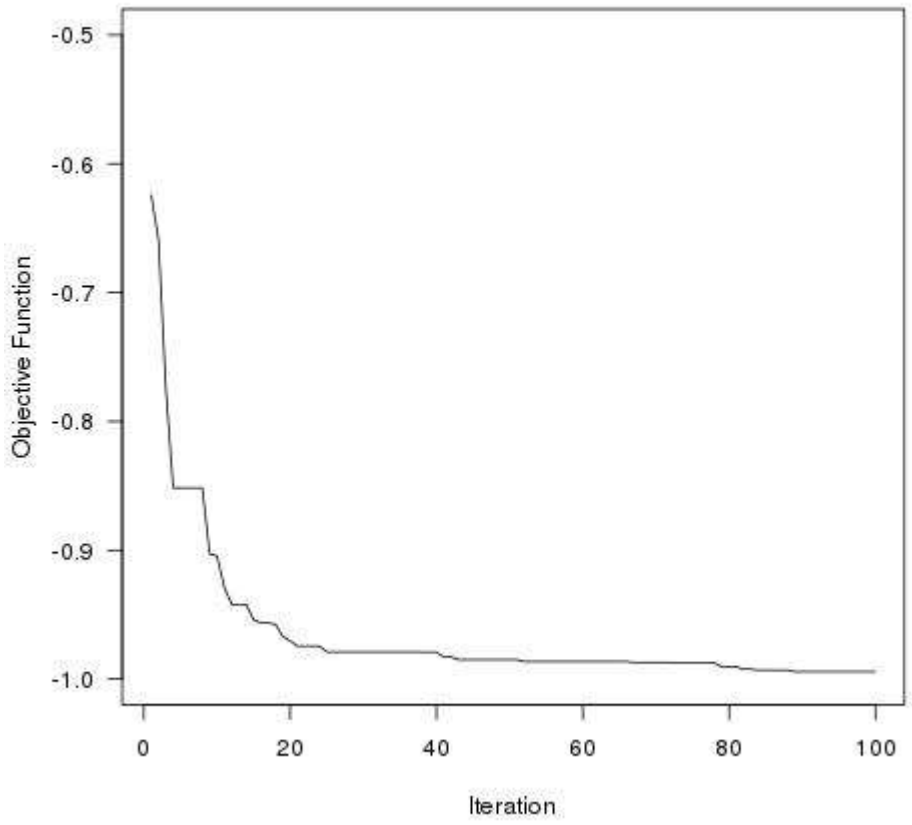
When the threshold accepting method was applied with the following parameters:

- the threshold values sequence: *ts* = 0.04, 0.04, 0.01,0.01
- number of iterations per threshold value: 25,

which means that there are 100 iterations in total, the correlation coefficient obtained was 0.975.

From the results obtained it can be observed that threshold accepting method did a slightly worse job in this case. Nevertheless, the results may be different if some other threshold sequence is chosen or if the number of iterations is higher (at the cost of a computation time).

Fig. 1: Local search



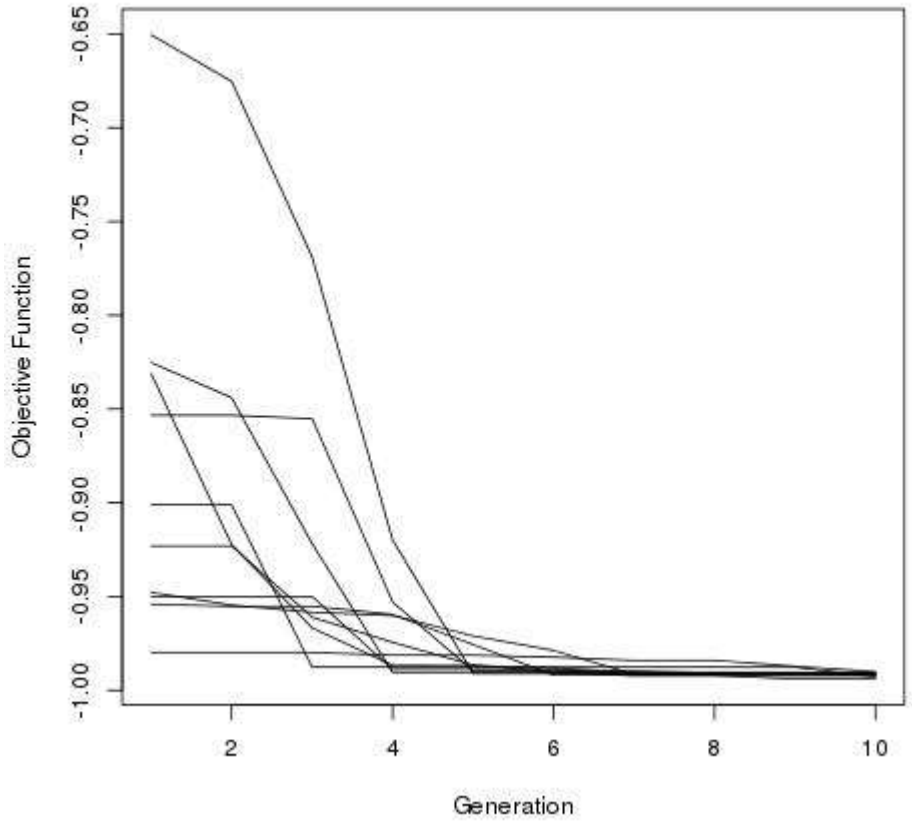
Source: own work

The application of the particle swarm optimization method with just 10 generations and the population size 10 (to make the 100 objective function evaluations in total, similarly as when using the previous methods) resulted in the correlation coefficient value 0.994 (see Figure 2, which shows how the objective function value changes with the iteration number for a sample run of the local search algorithm), which again may be considered a highly satisfactory result.

Conclusion

It was shown that the optimization problem considered, which was finding the optimal weights for particular variables for distance calculation using data on pairwise distances between subjects as input, can be efficiently solved using either the particle swarm optimization method or the local search method (or more generally the threshold accepting method). These rather general-purpose heuristic methods allow reaching a highly satisfactory value of the objective function and reach such solution quite quickly.

Fig. 2: Particle swarm optimization



Source: own work

Acknowledgment

This paper has been produced using the resources for the institutional support within the project IP 400040 at Faculty of Informatics and Statistics at University of Economics in Prague.

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. (2011). *Numerical methods and optimization in finance*. Waltham: Academic Press.

Gandomi, A., H. & Alavi, A., H. (2012). Krill herd: A new bio-inspired optimization algorithm. *Commun Nonlinear Sci Numer Simulat, 17.*

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27.

Kennedy, J. & Eberhart, R. (1995). Particle swarm optimisation. *Proc. of the IEEE Int .Conf. on Neural Networks,* Piscataway, NJ.

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

R Core Team (2014). 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* (4th ed.). New York: Springer.

Contact

Nikola Kaspříková

University of Economics in Prague, Faculty of Informatics and Statistics

Department of Mathematics

Nám.W.Churchilla 4

130 67 Praha 3

Czech Republic

Mail: nb33@tulipany.cz