Multiobjective clustering with automatic determination of the number of clusters

Julia Handl and Joshua Knowles

Keywords: evolutionary clustering, multiobjective clustering, evolutionary multiobjective clustering, determination of the number of clusters, data mining

Motivation:

Clustering methods, that is, algorithms for the identification of homogenous groups of data items within a data set, are fundamental tools in bioinformatics. The importance of such automated approaches to data processing and data analysis is further intensified by the scale of the data sets generated by today's high-throughput technologies; the applications of clustering methods in the field of post-genomics range from the identification of groups of coexpressed genes in microarray experiments to the identification of protein groups with structural and/or sequence similarity.

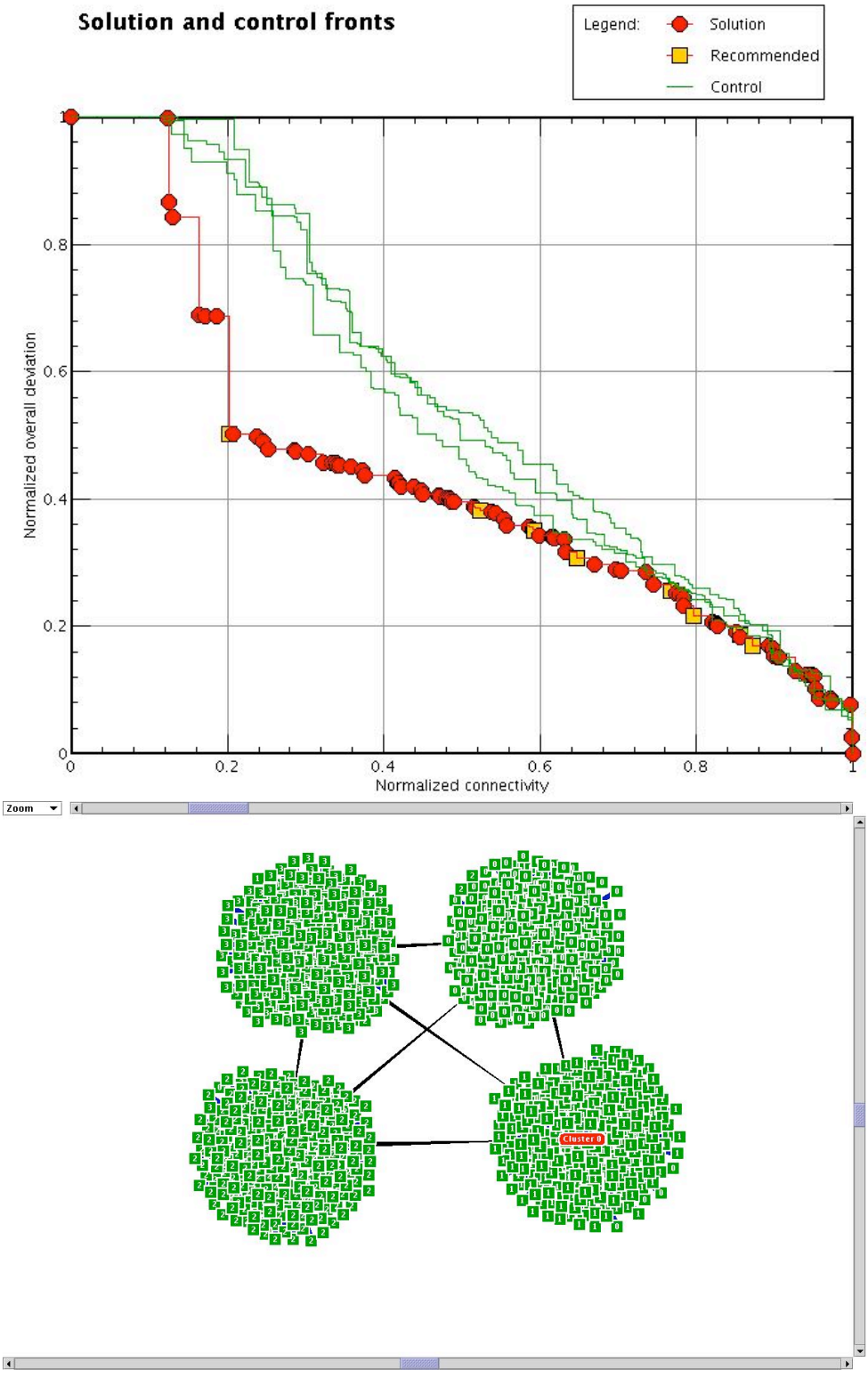
However, a fundamental dilemma in clustering is the difficulty of grasping the intuitive notion of a cluster by means of an explicit, formal

definition. While there are several valid properties that may generally be ascribed to a good clustering, these may be partially contradictory and/or inappropriate in a particular context. Yet, existing clustering methods attempt, explicitly or otherwise, to optimize just one such property; this confinement to a particular clustering criterion is one of the reasons for the fundamental discrepancies observable between the solutions produced by different algorithms and can cause a clustering method to fail in a context where the employed criterion is inappropriate.

In this work we show that the use of multiobjective optimization methods may provide a way out of this dilemma. We demonstrate that the simultaneous optimization of several clustering criteria significantly improves classification performance and results in an algorithm that is robust over a range of different data properties. In

Downloads:

- C++ source code and Java interface of the newest version of MOCK (see [2] below). Please drop me an email, if you download this software.
- Table providing Median and Interquartile Range values for the comparison of MOCK in [3] (against k-means, average link, single link and Strehl's ensemble method over 34 different data sets) (.ps file).
- Supporting material for VIENNA (data sets and additional results for [5]) (.zip file)
- Generators for the synthetic data sets used in [2] and [3].
- Generators for the more complex synthetic data sets used in [2].
- USM dissimilarity matrix for random computer files used in [1].



particular, the algorithm generates a set of trade-off solutions. These correspond to different compromises of the objectives employed, and thus provide a range of alternative hypotheses to the practitioner. Furthermore, the composition of the trade-off solutions may provide additional insight into the properties of the data, and thus increase confidence in the results obtained.

Publications:

1. Julia Handl and Joshua Knowles. (2005) Multiobjective clustering around medoids. Proceedings of the Congress on Evolutionary Computation (CEC 2005). Volume 1, pages 632--639. Copyright IEEE Press. PDF.

2. Julia Handl and Joshua Knowles. (2005) Improving the scalability of multiobjective clustering. Proceedings of the Congress on Evolutionary

Computation (CEC 2005). Volume 3, pages 2372--2379. Copyright IEEE Press. PDF.

3. Julia Handl and Joshua Knowles. (2005) Exploiting the trade-off -- the benefits of multiple objectives in data clustering. Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization (EMO 2005). Pages 547-560. LNCS 3410. PDF.

4. Julia Handl and Joshua Knowles. (2004) Multiobjective clustering with automatic determination of the number of clusters. Technical Report TR-COMPSYSBIO-2004-02. UMIST, Manchester, UK. PDF.

5. Julia Handl and Joshua Knowles. (2004) Evolutionary multiobjective

clustering. Proceedings of the Eighth International Conference on Parallel Problem Solving from Nature (PPSN VIII). Pages 1081-1091. LNCS 3242. Copyright Springer-Verlag. PDF.