Exploiting the Trade-Off — The Benefits of Multiple Objectives in Data Clustering

Julia Handl and Joshua Knowles

School of Chemistry, University of Manchester
Faraday Building, Sackville Street, PO Box 88, Manchester M60 1QD
http://dbk.ch.umist.ac.uk/handl/mock/

Abstract. In previous work, we have proposed a novel approach to data clustering based on the explicit optimization of a partitioning with respect to two complementary clustering objectives [6]. Here, we extend this idea by describing an advanced multiobjective clustering algorithm, MOCK, with the capacity to identify good solutions from the Pareto front, and to automatically determine the number of clusters in a data set. The algorithm has been subject to a thorough comparison with alternative clustering techniques and we briefly summarize these results. We then present investigations into the mechanisms at the heart of MOCK: we discuss a simple example demonstrating the synergistic effects at work in multiobjective clustering, which explain its superiority to single-objective clustering techniques, and we analyse how MOCK's Pareto fronts compare to the performance curves obtained by single-objective algorithms run with a range of different numbers of clusters specified.

Keywords: Clustering, multiobjective optimization, evolutionary algorithms, automatic determination of the number of clusters.

1 Introduction

Clustering is commonly defined as the task of finding natural groups within a data set such that data items within the same group are more similar than those within different groups. This is an intuitive but rather 'loose' concept, and it remains quite difficult to realize in general practice. Evidently, one reason for the difficulty is that, for many data sets, no unambiguous partitioning of the data exists, or can be established, even by humans. But even in cases where an unambiguous partitioning of the data *is* possible, clustering algorithms can drastically fail. This is because most existing clustering techniques rely on estimating the quality of a particular partitioning by means of just one *internal evaluation function*, an objective function that measures intrinsic properties of a partitioning, such as the spatial separation between clusters or the compactness of clusters. Hence, the internal evaluation function is assumed to reflect the quality of the partitioning reliably, an assumption that may be violated for certain data sets.

C. A. Coello Coello et al. (Eds.): EMO 2005, LNCS 3410, pp. 547–560, 2005.

Given that many objective functions for clustering are complementary, the simultaneous optimization of several such objectives may help to overcome this weakness. In previous work [6], we have demonstrated this idea, showing that the simultaneous optimization of two clustering objectives results in clear performance gains with respect to single-objective clustering algorithms. However, the algorithm presented, VIENNA (Voronoi initialized evolutionary nearest neighbour algorithm), was limited in two respects: its application required knowledge of the correct number of clusters, and no mechanism was presented to select good solutions from the Pareto front obtained.

Our new algorithm, MOCK (multiobjective clustering with automatic determination of the number of clusters), overcomes these weaknesses. It uses a novel, flexible representation that permits us to efficiently generate clustering solutions that both correspond to different trade-offs between our two clustering objectives and that contain different numbers of clusters. An automated technique is employed to select high-quality solutions from the resulting Pareto front, and it thus simultaneously determines the number of clusters in a data set. We briefly present the algorithm in this paper and summarize analytical results demonstrating its robust performance across data sets that exhibit a wide range of different data properties. A more detailed description and additional analytical results are provided in [7]. Besides introducing MOCK, a second goal of this paper is to give more insight into the mechanisms underlying multiobjective clustering: in particular, we aim to show that MOCK's good performance arises as a direct consequence of the simultaneous optimization of several clustering objectives: an archetypal problem — serving to illustrate the synergistic effects at work in multiobjective clustering — is used for this purpose. We additionally underline the differences between single- and multiobjective clustering by comparing the shape of MOCK's Pareto fronts to the performance curves obtained for single-objective clustering methods run with a varying number of clusters specified.

The remainder of this paper is organized as follows. Section 2 briefly summarizes related work on clustering and evolutionary algorithms. This is followed by a description of our algorithm, MOCK (Section 2.1), and all other contestant methods used in this study (Section 3). Section 4 presents our experiments and discusses results, and Section 5 concludes.

2 Related Work

Clustering problems arise in a variety of different disciplines, ranging from biology to sociology to computer science. Consequently, they have been the subject of active research for several decades, and a multitude of clustering methods exist nowadays, which fundamentally differ in their basic principles and in the properties of the data they can tackle. For an extensive survey of clustering problems and algorithms the reader is referred to Jain et al. [8].

Evolutionary algorithms (EAs) have a history of being applied to clustering problems [4, 12, 13]. However, previous research in this respect has been lim-