

Supporting Case-Based Retrieval by Similarity Skylines

Basic Concepts and Extensions

Eyke Hüllermeier, Ilya Vladimirskiy, Belén Prados Suárez, Eva Stauch

Conventional approaches to similarity search and case-based retrieval, such as nearest neighbor search, do require the specification of a global similarity measure which is typically expressed as an aggregation of local measures pertaining to different aspects of a case. Since the proper aggregation of local measures is often quite difficult, we propose a novel concept called *similarity skyline*. Roughly speaking, the similarity skyline of a case base is defined by the subset of cases that are most similar to a given query in a Pareto sense. Thus, the idea is to proceed from a d -dimensional comparison between cases in terms of d (local) distance measures and to identify those cases that are maximally similar in the sense of the Pareto dominance relation. To refine the retrieval result, we propose a method for computing maximally diverse subsets of a similarity skyline. Moreover, we propose a generalization of similarity skylines which is able to deal with uncertain data described in terms of interval or fuzzy attribute values. The method is motivated by and applied to similarity search over uncertain archaeological data.

1 Introduction

Similarity search in multi-dimensional data spaces is important for numerous application areas. In case-based reasoning (CBR), for example, it provides an essential means for implementing case retrieval, a critical step in case-based problem solving. In case-based retrieval, understood as the application of CBR paradigms to information retrieval tasks [2], similarity search becomes an even more central issue.

A commonly applied approach to case retrieval is the nearest neighbor (NN) search [3] which, despite its usefulness for certain problems, exhibits some disadvantages. Notably, NN methods assume a *global* similarity or, alternatively, distance function to be specified across the full feature set. The specification of such a measure often is greatly simplified by the “local–global principle”, according to which the global similarity between two cases can be obtained as an aggregation of various local measures pertaining to different features of a case [4]. However, even though it is true that local distances can often be defined in a relatively straightforward way, the *combination* of these distances can become quite difficult in practice, especially since different features may pertain to completely different aspects of a case. Moreover, the importance of a feature is often subjective and context-dependent. Thus it might be reasonable to free a user querying a system from the specification of an aggregation function, or at least to defer this step to a later stage.

In this paper, we propose a new concept, called *similarity skyline*, for supporting similarity search and case-based retrieval without the need to specify a global similarity measure. Roughly, the similarity skyline of a case base is defined by the subset of cases that are most similar to a given query in a Pareto sense [1]. More precisely, the idea is to proceed from a d -dimensional comparison between cases in terms of d (local) similarity or distance measures and to identify those

cases which are maximally similar in the sense of the Pareto dominance relation.

The next section describes the application that motivates our approach, namely similarity search in uncertain archaeological data. The concept of a similarity skyline is introduced in Section 3. In Section 4, we propose a method for refining the retrieval result, namely by selecting a (small) diverse subset of a similarity skyline. Section 5 is devoted to a generalization of similarity skylines which is able to deal with uncertain data described in terms of interval or fuzzy attribute values. Finally, Section 6 presents some experimental results, and Section 7 concludes the paper.

2 Motivation and Background

Even though the methods introduced in this paper are completely general, they have been especially motivated by an application in the field of archeology. As we shall report experimental results for this application later on, we devote this section to a brief introduction.

The DEADDY project aims at using knowledge discovery techniques to extract valuable information from archaeological databases. The domain under study is the analysis of graveyards in the Early Middle Ages. The data informs about graves, the persons buried therein, and the grave goods (objects which were put into the grave during the funeral ceremony according to religious rules or traditions typical for the given historical moment). Fig. 1 shows a screen shot of the DEADDY user interface. One can see a data record with information about particular grave goods: type, material, position in the grave, etc.

For experimental purpose (cf. Section 6), we have chosen the graveyard Wenigumstadt, which dates from the Early Middle Ages and is situated in the south of Germany. The inhabitants of a small village were buried in this cemetery from the end of the Roman Empire to the Age of Charle-