Information Systems 59 (2016) 2–14

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Efficient similarity search within user-specified projective subspaces

Michael E. Houle [a,*], Xiguo Ma [b], Vincent Oria [c], Jichao Sun [c]

[a] National Institute of Informatics, Tokyo 101-8430, Japan
[b] Google Mountain View, Mountain View, CA 94043, USA
[c] New Jersey Institute of Technology, Newark, NJ 07102, USA

ARTICLE INFO

Article history:
Received 31 January 2015
Received in revised form
1 December 2015
Accepted 19 January 2016
Available online 26 February 2016

Keywords:
Subspace similarity search
Multi-step search
Intrinsic dimensionality

ABSTRACT

Many applications — such as content-based image retrieval, subspace clustering, and feature selection — may benefit from efficient subspace similarity search. Given a query object, the goal of subspace similarity search is to retrieve the most similar objects from the database, where the similarity distance is defined over an arbitrary subset of dimensions (or features) — that is, an arbitrary axis-aligned projective subspace — specified along with the query. Though much effort has been spent on similarity search in fixed subspaces, relatively little attention has been given to the problem of similarity search when the dimensions are specified at query time. In this paper, we propose new methods for the subspace similarity search problem for real-valued data. Extensive experiments are provided showing very competitive performance relative to state-of-the-art solutions.

1. Introduction

Similarity search is of great importance to applications in many different areas, such as data mining, multimedia databases, information retrieval, statistics and pattern recognition. Specifically, a similarity query retrieves from the database those objects that most closely resemble a specified query object, based on some measure of pairwise similarity (typically in the form of a distance function). Due to its importance, much effort has been spent on the efficient support of similarity search. However, most existing approaches consider search only with respect to a fixed feature space. In this paper, we focus on the subspace similarity search problem, in which the calculation of similarity values is restricted to a subset of dimensions specified along with the query object. By specifying a subset of dimensions, the user indicates that the similarity computation should be computed as if the data had been projected to a desired axis-parallel subspace beforehand. The challenge is to efficiently preprocess the data in such a way that any such projective similarity query can be accommodated effectively.

As with similarity search on fixed spaces, subspace similarity search may also have an impact in application areas where the feature set under consideration changes from operation to operation. Such changes could be due to a modification of query preferences (as in content-based image retrieval), or to the determination of the local structure at different locations within data (as in subspace clustering), or to a systematic exploration of feature subspaces (as in feature selection). Motivated by the difficulty of search in higher dimensional spaces due to the so-called 'curse of dimensionality' [1–3], the efficiency of similarity search may be improved through an appropriate projection to a lower-dimensional subspace (as in dimensional reduction). In content-based image retrieval, images are often represented by feature vectors extracted based on

* Corresponding author.
E-mail addresses: meh@nii.ac.jp (M.E. Houle), maxiguo@google.com (X. Ma), oria@njit.edu (V. Oria), js87@njit.edu (J. Sun).

http://dx.doi.org/10.1016/j.is.2016.01.008
0306-4379/

color, shape, and texture descriptors. In an exploration of the dataset, a query involving one combination of features (such as color) may be followed by a query on a different combination (such as shape). In subspace clustering [4], the formation of an individual cluster is generally assessed with respect to a subset of features that most closely describe the concept associated with the cluster. Since verification of a cluster requires the identification of a feature subset together with an object subset, the effectiveness of the overall clustering process may depend on the efficient processing of subspace similarity queries. Wrapper methods for feature selection [5] require an evaluation process, such as k-nearest neighbor (k-NN) classification, for the identification of effective combinations of features. Exploration and visualization of feature subspaces [6,7] can be extremely time-consuming when the neighborhoods are determined using exhaustive search, due to the exponential number of potential combinations involved. To accelerate the process, the efficient support of subspace similarity search is needed.

Almost all existing similarity search indices require that the similarity measure and associated vector space both be specified before any preprocessing occurs. Since the subspaces to be searched are typically not known until query time, traditional methods for fixed spaces (as surveyed in [8]) cannot be effectively applied for the subspace search problem. Even if the query subspaces were known in advance, constructing an index for every possible subspace would still be prohibitively expensive. Of all the methods for similarity search appearing in the research literature, only very few have been specifically formulated for the subspace search problem; a survey of these methods will be presented in Section 2.1. In general, existing solutions for subspace similarity search suffer greatly in terms of the computational cost.

Of the two main types of similarity queries (k-NN queries and range queries), k-NN queries are often more important, due to the difficulty faced by the user in deciding range thresholds. This is especially the case for the search in subspaces, since the range values of interest will typically depend on the number of features associated with the subspace. In this paper, we focus only on k-NN queries.

We now formally define the subspace search problem for k-NN queries. Given an object domain $\mathcal{U}$, let $S \subseteq \mathcal{U}$ denote a set of database objects represented as feature vectors in $\mathbb{R}^D$. The set of features will be denoted simply as $F = \{1, 2, \ldots, D\}$, with feature $i \in F$ corresponding to the i-th coordinate in the vector representation. Let $d: \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ be a distance function defined for the vector space. Given an object vector $u = (u_1, \ldots, u_{|F|}) \in S$, its projection with respect to a feature subset $F' \subseteq F$ is the vector $u' = (u'_1, \ldots, u'_{|F|})$ such that for all $i \in F$, $u'_i = u_i$ whenever $i \in F'$, and $u'_i = 0$ otherwise. The feature set F' thus indicates a unique axis-aligned projective subspace to which distance calculations can be restricted.

Definition 1 (*Subspace k-NN query*). Given a query object $q \in \mathcal{U}$, a query subspace $F' \subseteq F$, and a query neighborhood size k, a subspace k-NN query $\langle q, F', k \rangle$ returns the k objects of S most similar to q, for the distance function $d_{F'}(q, u) \triangleq d(q', u')$, where q' and u' are the projections of q and u with respect to F', respectively.

Note that with this definition of the problem, the user is given complete freedom to specify any subspace with any number of dimensions. In particular, the number of possible choices of subspace in a D-dimensional full space is $2^D - 1$.

As an example of a subspace distance function, for any given $p \in [1, \infty)$, the L_p distance between two objects $q, u \in \mathcal{U}$ restricted to the axis-aligned projective subspace F' is defined as

$$d_{F'}(q, u) = \left(\sum_{i \in F'} |q_i - u_i|^p \right)^{\frac{1}{p}}.$$

Table 1 shows, for a set of 10 points in 5 dimensions, a 3-dimensional subspace similarity query result for the squared Euclidean distance.

The issues surrounding multidimensional query formulation, and in particular how users select data subspaces for exploration, are beyond the scope of this paper. Here, we limit our attention to the design of effective generic algorithms for the subspace similarity search problem. In particular, we follow the so-called 'multi-step' search strategy [9–11], utilizing 1-dimensional distances as lower bounds to efficiently prune the search space. The main contributions of this paper are:

- algorithms specifically tailored for both exact and approximate axis-parallel subspace similarity search in real-valued datasets, where the subspace dimensions are specified along with the query;

Table 1
Example of a subspace similarity query result for the choice of dimensions 2, 4, and 5, using the squared Euclidean distance.

Dimension	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	Variance	q
1	4.9	2.1	0.4	0.4	1.8	8.6	7.3	8.7	9.2	5.3	11.0481	9.0
2	5.5	4.4	6.8	8.6	0.5	9.8	9.8	5.8	2.9	0.5	10.6324	5.5
3	9.9	7.1	4.2	1.8	1.3	4.6	1.1	6.8	1.9	1.1	8.6016	1.0
4	9.3	8.2	7.3	6.4	8.3	4.3	2.3	1.8	1.1	3.2	8.2856	8.7
5	6.0	7.3	7.5	9.6	9.1	6.0	9.3	2.5	3.2	1.7	7.5496	6.5
Full *dist* & *rank* to q	96.63	86.92	88.85	99.11	83.85	51.22	58.39	97.43	76.26	91.99	–	0.0
	8	5	6	10	4	1	2	9	3	7	–	0
dist & *rank* to q in 2, 4, 5	0.61	2.10	4.65	24.51	31.92	38.10	55.49	63.70	75.41	78.29	–	0.0
	1	2	3	4	5	6	7	8	9	10	–	0