Information Sciences 179 (2009) 1234–1248

Contents lists available at ScienceDirect

Information Sciences

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

SubSpace Projection: A unified framework for a class of partition-based dimension reduction techniques

Hao Cheng [*], Khanh Vu, Kien A. Hua

School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Article history:
Received 20 January 2008
Received in revised form 30 April 2008
Accepted 1 December 2008

Keywords:
SubSpace Projection
Dimensionality reduction
Similarity search
Multidimensional indexing
Dimension partition

ABSTRACT

Similarity search in high dimensional space is a nontrivial problem due to the so-called curse of dimensionality. Recent techniques such as Piecewise Aggregate Approximation (PAA), Segmented Means (SMEAN) and Mean–Standard deviation (MS) prove to be very effective in reducing data dimensionality by partitioning dimensions into subsets and extracting aggregate values from each dimension subset. These partition-based techniques have many advantages including very efficient multi-phased approximation while being simple to implement. They, however, are not adaptive to the different characteristics of data in diverse applications.

We propose SubSpace Projection (SSP) as a unified framework for these partition-based techniques. SSP projects data onto subspaces and computes a fixed number of salient features with respect to a reference vector. A study of the relationships between query selectivity and the corresponding space partitioning schemes uncovers indicators that can be used to predict the performance of the partitioning configuration. Accordingly, we design a greedy algorithm to efficiently determine a good partitioning of the data dimensions. The results of our extensive experiments indicate that the proposed method consistently outperforms state-of-the-art techniques.

1. Introduction

Similarity search is expected to remain an active research area as efficient similarity query processing is of fundamental importance in various applications, such as time series [31], image retrieval [15] and text retrieval [35]. In these domains the exact matching may not always be preferable in different scenarios. Moreover, data collected are usually imperfect due to the effects of sampling, digitalization, lossy compressions and transformations [15]. Efficient similarity search, therefore, is important in these applications.

The goal of similarity search is to retrieve objects considered 'similar' to the object of interest within some user-specified threshold. There have been many models proposed to measure the (dis)similarity between objects, e.g., Euclidean distance (also known as the $\mathscr{L}_2$ norm) [33], general $\mathscr{L}_p$ norm [37], and Dynamic Time Warping [24]. In this paper, we adopt the $\mathscr{L}_2$ metric as it is the most common model and many other metrics are based on it [18,35]. Furthermore, it has been proved that *any finite metric space can be embedded into $\mathscr{L}_2$ norm space* [9].

In general, similarity retrieval requests are issued in the form of window queries, sphere queries, or k-nearest neighbor (kNN) queries. Window queries can be evaluated efficiently (e.g., as with the Pyramid method [8]), while kNN queries can be

[*] Corresponding author. Tel.: +1 407 823 3483; fax: +1 407 823 5419.
E-mail addresses: haocheng@eecs.ucf.edu (H. Cheng), khanh@eecs.ucf.edu (K. Vu), kienhua@eecs.ucf.edu (K.A. Hua).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.12.004

implemented on top of sphere queries [32]. In this paper, we focus on sphere queries in $\mathscr{L}_2$ norm space. The objects are denoted as vectors, e.g., $\vec{a} = (a_1, \ldots, a_m)$ in m-dimensional Euclidean space $\Re^m$.

Problem 1 (*Sphere Query*). Given a set of objects $A \in \Re^m$ and a query object $\vec{q} \in \Re^m$, find all objects $\vec{a} \in A$ whose $\mathscr{L}_2$ norm from $\vec{q}$ is no greater than a user-specified threshold ϵ, that is $\{\vec{a}|\vec{a} \in A,\ \mathscr{L}_2(\vec{a}, \vec{q}) \leqslant \epsilon\}$, in which $\mathscr{L}_2(\vec{a}, \vec{q}) = \left(\sum_{i=1}^m (a_i - q_i)^2\right)^{\frac{1}{2}}$.

When A is large and m is high, Problem 1 poses a serious challenge to efficient search of qualifying objects due to the so-called *curse of dimensionality* [30], which causes most indexing techniques to perform poorly [36]. To address this problem, recent techniques have been proposed to reduce dimensionalities while guaranteeing no-false-dismissal [18]. Among these, partition-based methods [22,37,33,34] divide dimensions into pre-determined disjoint groups and transform the original vectors into feature vectors with much lower dimensionality. Besides being easy to implement, these techniques are very competitive with more complex schemes. However, as the partitioning is fixed, they are not adaptive to the various characteristics of the data in diverse applications, and therefore may perform well on some datasets but not on others.

In this paper, we propose SubSpace Projection (SSP), a unified framework for the partition-based reduction techniques. In SSP, dimensions could be partitioned in any order and into groups with different sizes. We examine the effects of dimension partitioning on query performance, and show that the approximation performance can be predicted using parameters computed from the data. Following these findings, we devise a greedy algorithm to optimize the partitioning without the need to examine all possible partitions.

The rest of the paper is organized as follows. Section 2 provides a survey of recent techniques and an overview of our approach. The general framework of SSP is presented in Section 3. Section 4 explores configurations of SSP and the impact of dimension partitions. We solve the problem of finding a sub-optimal partition in Section 5 and report the results in Section 6. Finally we conclude the paper in Section 7.

2. Related work

Various dimension reduction methods have been proposed in the literature. Discrete Fourier Transform (DFT) [6] reduces dimensions by truncating data sequences by keeping only the low frequency Fourier coefficients. Discrete Wavelet Transform (DWT) [12] decomposes data sequences into the wavelet coefficients and discards those corresponding to higher resolutions. Singular Value Decomposition (SVD) [25] examines the characteristics of the dataset to find the optimal linear mapping of its data. It is well known that SVD has high computation overhead because of the eigen-decomposition and is not well suited to a dynamic database [21]. Orthogonal Locality Preserving Projection (OLPP) [10] derives a low-dimensional linear manifold embedding as the optimal approximation to the local neighborhood structures of the dataset [28], while it may neglect the global structures [13]. A recent proposal in [11] extracts principal coefficients from the data with regard to the Cheybshev polynomials.

Another popular reduction approach is to compute distances from each data point to some pre-selected reference points and in the querying phase, discard disqualifying points according to the triangle inequality rule of the metric space. For instance, OMNI-Family [16] picks a set of reference points such that they are farthest apart from each other. The distances from each data point to all the reference points are used in the filtering. Another distance-based technique, iDistance [20] builds the index based on the local structure of the data. Each data point is assigned to the closest pre-selected reference point and the distance between them is used in the pruning.

Recently, several partition-based methods were proposed, including Piecewise Aggregate Approximation (PAA) [22], Segmented Means (SMEAN) [37] and Mean–Standard deviation (MS) [33,34]. The dimensions, in their original order, are partitioned into disjoint subsets of equal size. In the Euclidean space, PAA and SMEAN are identical as they extract the *mean* of each vector's portion corresponding to each subset of the dimensions, whereas MS computes both the *mean* and the *standard deviation* (the definitions of mean and standard deviation will be provided in Section 4). These schemes are simple, easy to implement and yet outperform more sophisticated methods [33,34]. However as their mappings are static and not adaptive to the characteristics of the datasets being indexed, their performances vary greatly for different datasets. An improvement of PAA, APCA [23], enables an adaptive representation for each individual sequence by independently partitioning the dimensions into subsets of different sizes. Although effective in pruning power, this proposal is hard to implement in practice. APCA, however, did not consider partitioning dimensions in arbitrary orders. As shown later, this has a great effect on the query performance.

Recognizing the potentials of the partition-based approach in high dimensional search, we focus our attention to addressing the limitations of its representative techniques, namely expanding the capabilities of PAA, SMEAN and especially MS to deal with datasets with various characteristics. Specifically, we contribute the following:

(1) We propose a unified framework SSP and show that PAA, SMEAN and MS are instances of this class.
(2) We study the performance of query evaluations under various subspace selections (i.e. dimension partitions) and show that the summation of variances of features (*sv*) and its approximation (*sc*) are tell-tale indicators of the performance.
(3) We devise an efficient way to compute the approximate performance indicator (*sc*) without materialization of the partition and propose a greedy algorithm to efficiently derive a sub-optimal partition for a given dataset to achieve better query performance.