Information Sciences 281 (2014) 714–725

Contents lists available at ScienceDirect

Information Sciences

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

Nonnegative sparse locality preserving hashing

Cong Liu, Hefei Ling *, Fuhao Zou, Mudar Sarem, Lingyu Yan

School of Computer Science and Technology, Huazhong University of Science and Technology, 1037 Luoyu Rd., Hongshan, Wuhan 430074, Hubei, China

ARTICLE INFO

Article history:
Received 21 February 2013
Received in revised form 23 December 2013
Accepted 26 March 2014
Available online 2 April 2014

Keywords:
Approximate similarity search
Locality preserving projection
Sparse representation
Multiplicative updating
Nonnegative sparse locality preserving hashing

ABSTRACT

It is a NP-hard problem to optimize the objective function of hash-based similarity search algorithms, such as Spectral Hashing and Self-Taught Hashing. To make the problem solvable, existing methods have relaxed the constraints on hash codes from binary values (discrete) to real values (continuous). Then eigenvalue decomposition was employed to achieve the relaxed real solution. The main problem is that the signs of the relaxed continuous solution are mixed. Such results may deviate severely from the true solution, which has lead to significant semantic loss. Moreover, eigenvalue decomposition confronts singularity problem when the dimension of the data is larger than the sample size. To address these problems, we propose a novel method named Nonnegative Sparse Locality Preserving Hashing (NSLPH). Nonnegative and sparse constraints are imposed for a more accurate solution which preserves semantic information well. Then, we have applied nonnegative quadratic programming and multiplicative updating to solve the optimization problem, which successfully avoids the singularity problem of the eigenvalue decomposition. The extensive experiments presented in this paper demonstrate that the proposed approach outperforms the state-of-the-art algorithms.

1. Introduction

The similarity search is one of the most primitive operations in machine learning, pattern recognition, and information retrieval. It is also a crucial task in a lot of computer vision applications such as image clustering, object classification, multimedia or cross-media retrieval [1,2]. Recently, the efficiency of the similarity search algorithms becomes a critical issue because of the high dimensionality of the data which may incur "dimensionality curse". For objects in a low-dimensional feature space, the similarity search can be implemented efficiently with pre-build space-partitioning index structures (such as KD-tree [3]) or data-partitioning index structures (such as R-tree [4]). However, these methods are known to degrade dramatically to the brute-force query with the increment of the dimension in practice.

Recent hashing methods have been actively studied to provide efficient solution for high-dimensional data. These hashing techniques map similar objects into similar binary codes so that we can efficiently identify the approximate nearest neighbors by looking up similar binary codes based on the Hamming distance. Generally, the performance of the hash-based method significantly depends on the extent of the locality or the semantic preserving. That is to say, the hash codes of similar objects should be close to each other while those of dissimilar objects should be far apart after mapping. One of the most well-known hashing technique is the Locality Sensitive Hashing (LSH) [5]. Many researchers have proposed novel methods to improve the performance of LSH, such as Kernelized Locality Sensitive Hashing (KLSH) [6], and Coherency Sensitive

* Corresponding author. Tel.: +86 15007111150.
E-mail address: lhefei@hust.edu.cn (H. Ling).

http://dx.doi.org/10.1016/j.ins.2014.03.107
0020-0255/

Hashing (CSH) [8]. However, these methods were data-oblivious, and the design of the hash functions does not take the geometric structure of the objects into consideration, which might result in quite inefficient (or long) codes.

For the sake of more efficient codes, recent approaches endeavor to seek better data-aware hash functions via machine learning. The Spectral Hashing (SpH) [9] and the Self-Taught Hashing (STH) [10] are regarded as the state-of-the-art methods in this category. However, it is a NP-hard problem to optimize the objective function of these methods. To obtain an approximate solution, the constraints on the hash codes were relaxed from binary to real. Then, the Laplacian Eigenvalue decomposition (LapEig) [11] was employed to generate the relaxed real hash codes which deviate from the true binary solution. Thus, these methods encounter semantic loss issue resulting from the relaxition of the constraints. Moreover, these methods may confront singularity problem which results from the small sample problem. Some other excellent algorithms have been proposed in references [12–15]. However, these algorithms also relaxed the constraints on the hash codes from binary to real, which confront the same problem.

Recently, the sparsity and nonnegativity theory has been successfully used in the communities especially for face recognition and object clustering [16–21]. The Nonnegative Matrix Factorization (NMF) [16] could obtain a nonnegative part-based representation for each facial image, which was consistent with the cognition of the human brain. The implementation of the method showed that the nonnegative constraints have made the solution of the spectral clustering much closer to the ideal cluster indicator matrix [18]. The sparse constraints could well capture the intrinsic geometric (semantic) structure in the data [17]. That is to say, both constrains can alleviate semantic loss.

Motivated by the above observations, we have proposed a novel method named nonnegative sparse locality preserving hashing (NSLPH) which imposes the nonnegative and sparse constraints on Locality Preserving Projections (LPP)[22]. The LPP was applied instead of the LapEig because the LPP is a linear approximation of the LapEig, which could be used for out-of-sample data conveniently. The motivation of imposing nonnegative and sparse constraints is to make the relaxed solution closer to the true binary solution. This method has three contributions which can be highlighted as follows:

- It first introduces the sparse and nonnegative constraints into hashing. The nonnegative and sparse constraints insure that the solution is much closer to the ideal binary hash codes. Additionally, they can preserve intrinsic geometric structure, which can alleviate semantic loss.
- Because of the nonnegative constraints, the objective function could not be solved by the LapEig as it had been done in STH, SpH and LPP. Therefore, we have applied a novel method, the nonnegative quadratic programming [23] and multiplicative update [24], to solve the optimization problem. Our method is not only efficient, but it also successfully avoids the singularity problem which the LapEig might confront.
- In spite of locality preservation, we also consider the minimization of mapping error, which yields hash functions for efficient mapping on out-of -sample data, making it suitable for similarity search. Besides, our method generates hash function with good generalization ability, ensuring that the nearest neighbor search of the out-of-sample data in the high dimensional space will yield similar results in the low dimensional space.

The remainder of this paper is organized as follow. In Section 2, we discuss the related work focusing on the recent hashing methods and the sparse and nonnegative constraints. In Section 3, we present our method named Nonnegative Sparse Locality Preserving Hashing (NSLPH) in detail. The experimental results and analysis are demonstrated in Section 4. Finally, in Section 5, we make our conclusions.

2. Related works

2.1. Recent hashing methods

Recently, several novel methods which have attempted to learn hash functions for fast similarity search have been proposed. One of the most well-known hashing methods that preserve the similarity information is the Locality-Sensitive Hashing (LSH) [5]. In order to achieve better performance, the LSH method must increase the number of hash buckets and hash tables, which incurs more time and memory consumption. To overcome the drawback of the LSH method, the KLSH [6] and Multi-probe LSH [7] method were proposed. However, these methods are data-oblivious. The design of hash functions did not take the characters of the objects into consideration which might incur quite inefficient (or long) codes.

For the sake of more efficient codes, several recent approaches have attempted to seek better data-aware hash functions via machine learning. Most of these algorithms employed unsupervised learning to learn compact and effective hash codes[9,12,25–28]. Salakhutdinov et al. proposed the Semantic Hashing (SH) [25] method. The SH method applied Restricted Boltzmann Machines (RBM) to learn hash functions, and it has been shown that the learned hash codes preserve the semantic similarity in Hamming distance. The experiment results of the SH method illustrated that it was indeed able to generate compact and efficient binary codes and hence accelerated the similarity search. However, training the RBM was a computation-intensive and time-consuming process. This made the RBM expensive to retrain the hash function when the data evolved. Spectral Hashing (SpH), a brand-new method, applied the spectral graph theory to generate hashing code, and it could tackle the issue of data evolving. Although the SpH algorithm outperforms the SH algorithm, it has the drawback of impractical assumption. The SpH method assumed that the objects are spread in the Euclidean space with special distribution (either uniform distribution or Gaussian distribution) which was practically inapplicable.