Neurocomputing 167 (2015) 230–242

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised constraints preserving hashing

Di Wang, Xinbo Gao*, Xiumei Wang

VIPS Lab, School of Electronic Engineering, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 17 March 2015
Received in revised form
14 April 2015
Accepted 26 April 2015
Communicated by Luming Zhang
Available online 8 May 2015

Keywords:
Hashing
Binary codes
Semi-supervised learning
Nearest neighbor search
Pairwise similarity

ABSTRACT

With the ever-increasing amount of multimedia data on the web, hashing-based approximate nearest neighbor search methods have attracted significant attention due to its remarkable efficiency gains and storage reductions. Traditional unsupervised hashing methods are designed for preserving distance metric similarity which may lead to semantic gap among the high-level semantic similarities. Recently, attentions have been paid to semi-supervised hashing methods which can preserve data's a few available semantic similarities (usually given in terms of labels, pairwise constraints, tags, etc.). However, these methods often preserve semantic similarities for low-dimensional embeddings. When converting low-dimensional embeddings into binary codes, the quantization error will be accumulated thus resulting in performance deterioration. To this end, we propose a novel semi-supervised hashing method which preserves pairwise constraints for both low-dimensional embeddings and binary codes. It first represents data points by cluster centers to preserve data neighborhood structure and reduce the dimensionality. Then the constraint information is fully utilized to embed the derived data representations into a discriminative low-dimensional space by maximizing discriminative Hamming distance and data variance. After that, optimal binary codes are obtained by further preserving the semantic similarities in the process of quantizing the low-dimensional embeddings. By utilizing constraint information in the quantization process, the proposed method can fully preserve pairwise semantic similarities for binary codes thus leading to better retrieval performance. Thorough experiments on standard databases show the superior performance of the proposed method.

1. Introduction

With the explosive growth of data on the web, there is an urgent demand of approximate nearest neighbor (ANN) search methods to efficiently exploit user intent information from large web databases. The main challenges for ANN methods are fast query and low storage requirement. To meet this goal, various hashing methods have been proposed recently. Hashing aims to map high-dimensional data, such as documents, images, or videos, into a set of low-dimensional compact binary codes while preserving the underlying similarity relationship of the original data. The pairwise similarity comparisons between these binary codes can be measured by Hamming distances which only involve efficient bit-count operations thus can be computed very quickly. Furthermore, only a small number of bits are sufficient for binary codes to maintain the information required for retrieval, which bring enormous storage savings. Due to these merits, hashing methods have been successfully used in various applications such as large-scale retrieval [1–3], feature descriptor learning [4,5], and near-duplicate detection [6–8].

According to how much supervised information is used, hashing methods can be classified as unsupervised, semi-supervised, and supervised methods. Most representative hashing methods are unsupervised, which are seeking to employ data information to compute binary codes. Early endeavors in unsupervised hashing concentrated on data-independent methods which are using random projections to construct hash functions without exploiting the knowledge of training data. Notable examples include Locality-Sensitive Hashing (LSH) [9], Kernelized LSH (KLSH) [10], and Shift-Invariant KLSH (SKLSH) [11]. However, due to the limitation of random hash functions generating scheme, LSH-like hashing methods need long binary codes to achieve reasonable performance, therefore suffer from long query time and high storage cost. To improve the limitation of data-independent method, recent research focuses on data-dependent hashing methods which can generate hash functions by making use of the correlations between data points. Typical approaches are PCA-like hashing [12–15], manifold-like hashing [16–19], and *K*-means-like hashing [20–22].

To preserve the semantic similarities, supervised hashing algorithms have been employed to design more effective hash functions. Linear discriminant analysis hashing (LDAHash) obtains hash functions

* Corresponding author.
E-mail address: xbgao.xidian@gmail.com (X. Gao).

http://dx.doi.org/10.1016/j.neucom.2015.04.072
0925-2312/

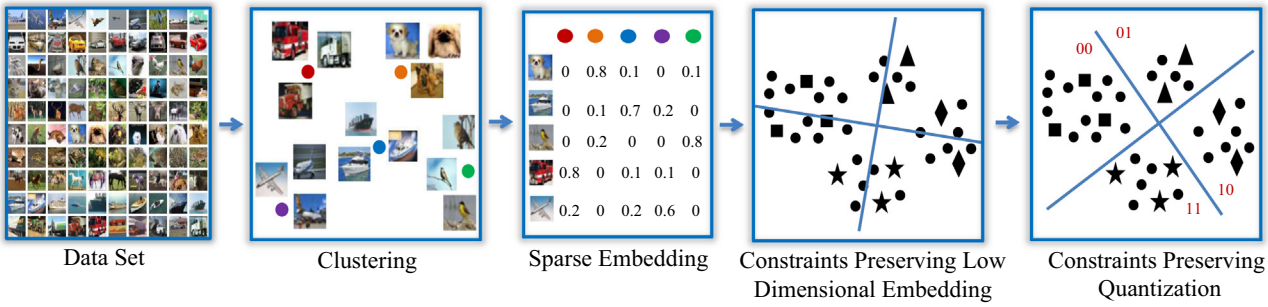

Fig. 1. Flowchart of the proposed semi-supervised constraints preserving hashing method.

by maximizing the between-class scatter among binary codes associated with different classes [4]. Binary reconstructive embedding (BRE) learns hash functions by minimizing the reconstruction error between the original semantic similarities and the Hamming distances of the corresponding binary embeddings [23]. Minimal loss hashing (MLH) introduces a hinge-like loss function depending on semantic similarity information and learns binary codes based on structural prediction [24]. Supervised hashing with kernels (KSH) maps data to hash codes whose Hamming distances are minimized on similar pairs and simultaneously maximized on dissimilar pairs [25]. Although the LDAHash can tackle supervision via easy optimization, it lacks of adequate performance. BRE, MLH and KSH can obtain better search accuracy than LDAHash whereas they often attribute to sophisticated optimization and have expensive training cost. This imperfection has greatly diminished the applicability of these methods to large-scale tasks.

Usually, labeling samples requires much human expertise on large scale data set. For this reason, available supervised information can be very limited. Semi-supervised methods will be helpful in this case. They often take advantage of both supervised information and data's underlying similarity information. Label-regularized max-margin partition (LAMP) method uses kernel-based similarity and additional small number of pairwise constraints as side information to generate hash codes [26]. The encoding process is formulated within a regularized maximum margin framework which can be solved by a series of convex sub-problems with quadratic-program. However, the optimization process is so complex that leading lengthy training process and reducing the serviceability. Semi-supervised hashing (SSH) minimizes the empirical error on the labeled data while maximizing entropy of hash bits over the labeled and unlabeled data [27]. According to different optimization algorithms, SSH has three solutions, SSH-Orthogonal, SSH-Nonorthogonal and sequential projection learning for hashing (SPLH). SSH-Orthogonal and SSH-Nonorthogonal first project the original data into low-dimensional embeddings, and then quantize the embeddings into binary ones by thresholding. However, the quantization error produced by this conversion process will decrease the hashing performance. To reduce the quantization error, SPLH is designed as a boosting learning method. Each hash function in SPLH is intended to correct the errors produced by the previous ones. The hash functions are learned iteratively such that the pairwise label matrix is updated by imposing higher weights on point pairs violated the preceding hash function. However, SPLH judges every previous bit separately when deciding the errors of the obtained projections to penalize them with the higher weights. Since the similarities of hash codes should take consideration of all bits holistically, SPLH may incur more errors. To solve this problem, Bootstrap sequential projection learning for semi-supervised nonlinear hashing (Bootstrap-NSPLH) method is proposed [28]. It utilizes bootstrap sequential projection learning to rectify the quantization errors by taking into consideration of all the previously learned bits holistically. However, Bootstrap-NSPLH is mainly designed for label information. When weaker supervised information such as pairwise constraint is available, it needs calculating the pairwise Hamming similarity matrix (for N training data points, the size of this matrix will be $N \times N$) for the whole data set as pairwise constraint information is often scattered. This is intolerable for the large data set for the enormous storage space and large amount of calculation.

To effectively reduce the quantization error accumulated during converting low-dimensional embeddings into binary codes after relaxation and efficiently use the pairwise constraints, we propose semi-supervised constraints preserving hashing (SCPH) method in this paper. Fig. 1 illustrates the flowchart of the proposed method. It first partitions the data points into some clusters and represents the data points by cluster centers to preserve data neighborhood structure and reduce dimensionality. Then constrain information is fully utilized to embed the derived data representations into a discriminative low-dimensional space by maximizing discriminative distance and data variance. After that, optimal binary codes are obtained by preserving the semantic similarities in the process of quantizing the low-dimensional embedding. The main contributions are summarized as follows:

- A discriminative low-dimensional space is learned in which points with similar constraints are pushed as close as possible, while points with dissimilar constraints are pulled away as far as possible.
- Constraint information is further utilized to guide the quantization process. The obtained discriminative low-dimensional embeddings are quantized into Hamming space by jointly maximizing the binary codes' discriminability and minimizing the quantization error. This makes the similarities of binary codes consistent with the pairwise constraints as much as possible.
- The proposed method has more applicability as the weaker supervised information that is pairwise constraint can be used. When other supervised information for example label or tag is available, it is also useful as pairwise constraint can be readily derived from them.

The rest of this paper is organized as follows. The proposed SCPH method is presented in Section 2. Section 3 gives its computational complexity analysis. Section 4 shows the experimental results and analysis on popular image data sets. Finally, Section 5 is the conclusion.

2. Semi-supervised constraints preserving hash codes learning

Given a training set of n data points, denoted as $\{x_1, x_2, \ldots, x_n\}$ with $x_i \in \mathbb{R}^d$, which form the rows of the data matrix $X \in \mathbb{R}^{n \times d}$. The data set also includes a fraction of pairwise constraint information with similar pair set M in which $(x_i, x_j) \in M$ implies that x_i and x_j belong to the same class, and dissimilar pair set C in which $(x_i, x_j) \in C$ implies that x_i and x_j belong to different classes. Our goal is to learn binary codes $Y \in \{1, -1\}^{n \times r}$ [1] of the data set through

[1] Converting $-1/1$ codes to 0/1 codes is a trivial shift and scaling operation.