UDC 004.021

INFLUENCE OF DIMENSIONALITY REDUCTION METHODS ON LSH PERFORMANCE IN CONTEXT OF DIMENSIONALITY CURSE MITIGATION

Shmalko F.A.
National University "Zaporizhzhia Polytechnic"
Zaporizhzhia, Ukraine
https://orcid.org/0009-0006-0651-6448; shmalko.fedor@gmail.com

Abstract. *The paper analyzed the impact of the dimensionality reduction methods t-SNE and UMAP on the efficiency of Locality-Sensitive Hashing when searching for approximate nearest neighbors in high-dimensional spaces. It was also shown that the use of these methods increases data density, improves hashing quality, and reduces computational costs. The results obtained confirm the feasibility of their use in systems for processing large multidimensional data, in particular in clustering, information retrieval, and machine learning tasks.*

Keywords*: Locality-Sensitive Hashing; dimensionality reduction; t-SNE; UMAP; curse of dimensionality.*

Introduction. Locality-Sensitive Hashing (LSH) is an effective method for approximate nearest neighbors (ANN) search, that implies the utilization of hash-functions when grouping similar objects in the same buckets. However, its central advantage lies within the fact that it is able to be used in high-dimensional spaces that are generally subjected to the problem known as “curse of dimensionality” [1]. This phenomenon occurs when the number of space dimensions increases, so the noticeable part of data processing algorithms become either less effective or lose their discriminatory ability, because both the data becomes sparse due to the fact that data points are located so far from one another, that even the nearest neighbors can have relatively high distances – so the proximity becomes less evident; and also due to the fact that the high number of dimensions causes the distribution of even distances between data points to be almost the same, which means that the difference between the nearest and the furthest neighbors becomes insignificant, hindering the clustering and ANN search. The usage of dimensionality reduction methods, like UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-Distributed Stochastic Neighbor Embedding) removes these problems, because it allows the output high-dimensional data to be projected into the lower-dimensional space, preserving the data points semantic proximity, which provides effective ANN search with LSH, because hashing now occurs in a more compact and coherent space [2].

The research aims to evaluate how dimensionality reduction methods such as UMAP and t-SNE improve the performance of Locality-Sensitive Hashing (LSH) under conditions affected by the curse of dimensionality. The study focuses on analyzing their impact on data density, clustering precision, and computational efficiency in high-dimensional spaces.

The main part. It’s important to highlight that the increase in space dimensionality provokes further sparsification of data distribution, which complicates similar objects identification. If the output high-dimensional space is denoted as $\mathbb{R}^d$, then under the conditions of high dimensionality d, the ratio between the distances of arbitrary pairs of points becomes almost the same:

$$\frac{\|x_i - x_j\|}{\max\|x_k - x_l\|} \to 1, \quad \forall_{x_i, x_i, x_i, x_i} \in \mathbb{R}^d. \tag{1}$$

In such a situation, hash functions of LSH that rely on relative distances between objects, lose their efficiency because the probability of similar objects collision becomes low. This leads to the necessity of either increasing the hash-functions k, or creating more hash-tables L, which translates into higher computational overhead

Dimensionality reduction methods remove this problem, projecting points into a lower dimensional space $\mathbb{R}^{d'}$, where $d' \ll d$, but at the same time retaining important local and global relationships between the points.

t-SNE method performs non-linear reduction of dimensionality by iteratively approximating the distribution of pair points in a lower-dimensional space. Let p_{ij} be the probability of two points` x_i and x_j proximity in the output space:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_{k\neq l}\exp(-\|x_k - x_l\|^2/2\sigma^2)}. \quad (2)$$

In a lower-dimensional space the analogous distribution q_{ij} is built, which is defined by Student function with one degree of freedom:

$$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k - y_l\|^2)^{-1}}. \quad (3)$$

The t-SNE cost function that has to be minimized is defined as a Kullback-Leibler divergence between the p_{ij} and q_{ij}distributions:

$$C = \sum_{i\neq j} p_{ij} \log\frac{p_{ij}}{q_{ij}}. \quad (4)$$

So the usage of this method allows similar objects to be grouped in dense local regions, which makes the further hashing more efficient, because LSH performs better in case of data points being evenly distributed in a compact space.

UMAP, on the other hand, is a method which bases itself on topological analysis of local relationships between data points [3]. At first, it builds a weighted graph C, where weights are defined by the probability of two points x_i, x_j being connected:

$$w_{ij} = \exp(-d(x_i, x_j)/\rho_i). \quad (5)$$

Thereafter the optimization is performed through maximizing the preservations of relationships in a lower-dimensional space, using the sigmoid function of probabilities distribution:

$$q_{ij} = \frac{1}{1+a\|y_i - y_j\|^{2b}}. \quad (6)$$

The cost function then looks like:

$$C = \sum_{i\neq j} w_{ij} \log\frac{w_{ij}}{q_{ij}}. \quad (7)$$

After integrating either t-SNE or UMAP, the output space $\mathbb{R}^d$ is projected in the space $\mathbb{R}^{d'}$, where d' is chosen so that the sufficient informativeness of data is preserved, while minimizing computational overhead. Then classical hash-functions of LSH are performed:

$$h(x) = \lfloor a \cdot x + b \rfloor \mod w, \quad (8)$$

where $a \sim N(0,1)$, $b \sim U(0,w)$, and w is a bucket width.

Conclusion. Dimensionality reduction methods t-SNE and UMAP play a key role in drastically increasing LSH efficiency in high dimensional spaces. They help mitigate the “curse of dimensionality”, increase the local data density and decrease computation costs during ANN search. The usage of these methods in structured hashing allows for significant efficiency improvement of systems that are specialized on processing big multimodal data volumes.

REFERENCES

[1] Ghojogh, B., Crowley, M., Karray, F., & Ghodsi, A. (2023). *Elements of Dimensionality Reduction and Manifold Learning*. Springer. https://doi.org/10.1007/978-3-031-10602-6

[2] Damrich, S., Böhm, J., Hamprecht, F. & Kobak, D. (2022). "*Contrastive Learning Unifies t-SNE and UMAP*" arXiv. https://doi.org/10.48550/arXiv.2206.01816.

[3] McInnes L., Healy J. & Melville J. (2018). "*UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*" arXiv. https://doi.org/10.48550/arXiv.1802.03426.