REVIEW:Clustering high dimensional data,Ira Assent⁎

July 19, 2023

Curse of Dimensionality and Algorithmic Distortion

Unresolved Research Questions:

How can we effectively identify critical features or dimensions in high-dimensional datasets to reduce the impact of the curse of dimensionality on clustering performance?

What are the most appropriate distance metrics or similarity measures for measuring similarity between data points in high-dimensional spaces?

How can we develop robust and efficient clustering algorithms that can handle high-dimensional data and alleviate the curse of dimensionality?

What techniques can be employed to assess the quality and interpretability of clustering results in high-dimensional data when traditional visualization methods become less informative?

How does the curse of dimensionality affect the computational complexity of clustering algorithms, and how can we optimize algorithms to handle large-scale high-dimensional datasets?

Curse of Dimensionality and Algorithmic Distortion:

The "Curse of Dimensionality" is a well-known challenge in data mining and machine learning, which arises when dealing with high-dimensional datasets. As the number of input variables, or features, increases, the data points become increasingly sparse in the high-dimensional space. This sparsity leads to various issues, often referred to as algorithmic distortion, which can significantly impact the performance of clustering algorithms.

Increased Sparsity: With higher dimensions, data points are distributed farther apart from each other, making it difficult to identify meaningful patterns or clusters. Clusters become less distinguishable, leading to reduced accuracy in clustering results.

Diminished Distance Measures: Traditional distance metrics like Euclidean distance may lose their effectiveness in high-dimensional spaces due to the "crowding" effect, where data points tend to be equidistant from each other, resulting in an inadequate representation of their actual similarity.

Overfitting and Noise Sensitivity: High-dimensional data is prone to overfitting, where clustering algorithms may produce clusters that are more reflective of noise than genuine underlying patterns. This is especially problematic when there are more dimensions than relevant features.

Computationally Demanding: The increased dimensionality leads to higher computational complexity in clustering algorithms, making them computationally demanding and often impractical for large-scale datasets.

To address the curse of dimensionality, researchers explore various approaches, such as dimensionality reduction techniques like PCA or t-SNE, feature selection, or employing specialized clustering algorithms designed to handle high-dimensional data effectively.

In conclusion, the curse of dimensionality poses significant challenges in clustering high-dimensional data, resulting in algorithmic distortion that impacts clustering accuracy, interpretability, and computational efficiency. Unresolved research questions revolve around developing effective techniques to mitigate these challenges, enhance clustering performance, and gain valuable insights from complex high-dimensional datasets.

REFERENCE: Clustering high dimensional data,Ira Assent∗

To leave a comment, click the button below to sign in with Google.

RISHABH JAGTAP

Archive