Normalized Geometric Mean Distance

Kenneth Zhang

March 31, 2023

Abstract

The NGMD metric provides a measure of the quality of clustering by calculating the geometric mean of the pairwise distances between the centroids of the clusters, normalized by the distance between the two farthest points in the dataset. This metric is useful in determining how well the clustering algorithm has grouped similar data points together while keeping the clusters well separated from each other. A lower NGMD value indicates that the clustering algorithm has produced a better clustering solution. Therefore, NGMD can be an important tool in evaluating clustering algorithms and comparing different clustering solutions.

1 Introduction

The NGMD (Normalized Geometric Mean Distance) metric is a measure of how well a clustering algorithm partitions a dataset into distinct groups. The metric calculates the geometric mean of the pairwise distances between the centroids of the clusters, and then normalizes this value by the distance between the two farthest points in the dataset. This normalization step ensures that the NGMD metric always has a value between 0 and 1, where a value of 1 indicates perfect clustering and a value of 0 indicates random clustering. In this proof, we provide the mathematical formula for calculating the NGMD metric, including the steps for computing the centroids and pairwise distances between them.

2 Main Result

The main result of the proof is a mathematical formula for the NGMD metric, which can be used to evaluate the quality of clustering algorithms. The formula involves calculating the pairwise distances between the centroids of the clusters and normalizing this value by the distance between the two farthest points in the dataset. The resulting value ranges from 0 to 1, where a value of 1 indicates perfect clustering and a value of 0 indicates random clustering. This formula provides a quantitative measure of the clustering quality and can be used to compare the performance of different clustering algorithms.

3 Mathematic Intuition and Proof

Let X be a dataset of n points in d-dimensional space and let $C = C_1, C_2, \ldots C_k$ be a clustering of X into k clusters. The NGMD metric is defined as follows:

First, the centroids of the k clusters are computed:
$c_i = (1/|C_i|) * \sum(x \in C_i)x$, where $|C_i|$ is the number of points in cluster C_i and $sum(x \in C_i)$ x is the sum of all points in cluster C_i.

Next, the pairwise distances between the centroids are calculated:
$d_ij = ||c_i - c_j||$ where $||.||$ denotes the Euclidean distance between two points.

The geometric mean of these pairwise distances is then computed:
$G = (\prod_{i<j} d_ij)^{1/\binom{k}{2}}$, where $\binom{k}{2}$ is the binomial coefficient that counts the number of pairwise distances between the k centroids. Finally, the NGMD metric is obtained by normalizing the geometric mean by the distance between the two farthest points in X: $NGMD = G/max_{x,y \in X}||x - y||$, where $max_{x,y \in X}||x - y||$ is the maximum pairwise distance between any two points in our dataset, X.

This normalization step ensures that the NGMD metric is always between 0 and 1, where a value of 1 indicates perfect clustering and a value of 0 indicates random clustering.

4 Conclusion

In conclusion, the NGMD metric provides a valuable tool for evaluating the performance of clustering algorithms. By computing the pairwise distances between the centroids of the clusters and normalizing this value by the distance between the two farthest points in the dataset, the NGMD metric can provide a quantitative measure of the clustering quality. This metric can be used to compare the performance of different clustering algorithms and can help guide the selection of the optimal number of clusters. Overall, the NGMD metric is a powerful tool for assessing the quality of clustering algorithms and can play a key role in the development of effective data analysis strategies.