International Journal of Grid and High Performance Computing
Volume 16 • Issue 1

Pre-Cutoff Value Calculation Method for Accelerating Metric Space Outlier Detection

Honglong Xu, Foshan University, China
https://orcid.org/0000-0002-8645-9028
Zhonghao Liang, Foshan University, China
Kaide Huang, Foshan University, China*
Guoshun Huang, Foshan University, China
Yan He, Foshan University, China

ABSTRACT

Outlier detection is an important data mining technique. In this article, the triangle inequality of distances is leveraged to design a pre-cutoff value (PCV) algorithm that calculates the outlier degree pre-threshold without additional distance computations. This algorithm is suitable for accelerating various metric space outlier detection algorithms. Experimental results on multiple real datasets demonstrate that the PCV algorithm reduces the runtime and number of distance computations for the iORCA algorithm by 14.59% and 15.73%, respectively. Even compared to the new high-performance algorithm ADPOD, the PCV algorithm achieves 1.41% and 0.45% reductions. Notably, the non-outlier exclusion for the first data block in the dataset is significantly improved, with an exclusion rate of up to 36.5%, leading to a 23.54% reduction in detection time for that data block. While demonstrating excellent results, the PCV algorithm maintains the data type generality of metric space algorithms.

KEYWORDS

distance triangle inequality, index, metric space, outlier detection, pre-cutoff value

INTRODUCTION

The continuous expansion of data volumes and innovative applications propels the burgeoning growth of the big data industry. The cumulative volume of data places significant stress on data storage capabilities. However, simply increasing storage devices is not a sustainable solution. Viable strategies include timely data analysis and mining, as they alleviate the pressure on storing raw data and maximize data value.

As a core step in data processing, data mining is an active academic research field, giving rise to various techniques, such as classification, clustering, association analysis, and outlier detection. While most data mining techniques focus on discovering regular patterns within datasets, non-regular

DOI: 10.4018/IJGHPC.334125 *Corresponding Author

patterns are sometimes equally valuable. Outlier detection algorithms constitute a vital branch of data mining designed to identify these non-regular patterns. They enable the discovery of distinctive data points within a dataset, known as outliers or anomalies. Thanks to outlier detection, we can promptly identify exceptional events within data, such as network attacks (Catillo et al., 2023), fraudulent transactions (Hilal et al., 2022), or equipment malfunctions (Nesa et al., 2018), thus reducing losses. Moreover, outlier detection contributes to enhancing data quality (Larson et al., 2019), recognizing exceptional outcomes within datasets, and even unearthing new knowledge.

Research into outlier detection algorithms focuses on either specialized or generic algorithms. Specialized algorithms are tailored and optimized for the characteristics of data in various domains, making full use of available information to expedite outlier detection. However, in the era of big data, the ever-expanding and diverse data types pose significant challenges to the design capacity of specialized algorithms. In such circumstances, generic outlier detection algorithms have emerged. Generic algorithms abstract and unify commonalities across data types in different domains, performing searches and mining using only partial information from the dataset, thereby enabling the application of a single algorithm across diverse data types. While there may be some performance trade-offs, generic algorithms significantly reduce data mining systems' development and maintenance costs, bringing them significant attention in academic and industrial circles in recent years.

Fortunately, most data types can be designed with distance functions that adhere to the triangle inequality, allowing them to be mapped to metric spaces (Mao et al., 2015). This, in turn, facilitates the use of metric space algorithms for retrieval, analysis, and mining. Metric space outlier detection utilizes a definition of outliers entirely based on distance and detection algorithms that are also fully reliant on distance. It eschews the use of any information other than distance, making it applicable to a wide range of data types.

Metric space outlier detection algorithms belong to the distance-based outlier detection algorithm class. Much like their counterparts, they employ distance triangle inequality for pruning to eliminate non-outliers and accelerate the outlier detection process. This step heavily depends on the cutoff value of the outlier degree, where a higher value leads to improved efficiency in non-outlier exclusion. However, existing metric space outlier detection algorithms face a critical issue when detecting the first data block, where the available outlier degree cutoff value is set to 0, causing significant delays in detecting that block and severely impeding overall detection efficiency. To address this problem, we propose a pre-cutoff value calculation method that can be used to accelerate metric space outlier detection. By making full use of the distance triangle inequality, this method calculates an initial outlier degree cutoff value, hereafter referred to as the "pre-cutoff value," through minimal computations.

The main contributions of this paper are summarized as follows:

(1) Analyzing the time allocation of each data block in the outlier detection process reveals that the time cost of the first data block is significantly greater than that of other blocks, and we analyze the reasons behind this.
(2) Introducing a pre-cutoff value calculation method to accelerate metric space outlier detection, utilizing the distance triangle inequality, and designing a method that does not require additional distance computations.
(3) Proving mathematically that the pre-cutoff value–based accelerated outlier detection algorithm guarantees correct results.

The subsequent sections of this paper are outlined as follows: The next section provides an overview of distance-based outlier detection algorithms, including some metric space outlier detection algorithms. After that, we introduce the pre-cutoff value calculation method for accelerating metric space outlier detection and then analyze and prove its correctness. The following section presents the experimental results and provides an analysis. Finally, the last section summarizes the paper's work and outlines potential future work.