Information Systems 55 (2016) 37–53

Contents lists available at ScienceDirect

Information Systems

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

Efficient and flexible algorithms for monitoring distance-based outliers over data streams

Maria Kontaki, Anastasios Gounaris*, Apostolos N. Papadopoulos, Kostas Tsichlas, Yannis Manolopoulos

Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 1 April 2014
Received in revised form
22 April 2015
Accepted 7 July 2015
Recommended by: Martin Theobald
Available online 4 August 2015

Keywords:
Stream data mining
Outlier detection

ABSTRACT

Anomaly detection is considered an important data mining task, aiming at the discovery of elements (known as outliers) that show significant diversion from the expected case. More specifically, given a set of objects the problem is to return the suspicious objects that deviate significantly from the typical behavior. As in the case of clustering, the application of different criteria leads to different definitions for an outlier. In this work, we focus on distance-based outliers: an object x is an outlier if there are less than k objects lying at distance at most R from x. The problem offers significant challenges when a stream-based environment is considered, where data arrive continuously and outliers must be detected on-the-fly. There are a few research works studying the problem of continuous outlier detection. However, none of these proposals meets the requirements of modern stream-based applications for the following reasons: (i) they demand a significant storage overhead, (ii) their efficiency is limited and (iii) they lack flexibility in the sense that they assume a single configuration of the k and R parameters. In this work, we propose new algorithms for continuous outlier monitoring in data streams, based on sliding windows. Our techniques are able to reduce the required storage overhead, are more efficient than previously proposed techniques and offer significant flexibility with regard to the input parameters. Experiments performed on real-life and synthetic data sets verify our theoretical study.

1. Introduction

Anomaly detection is a data mining task focusing on the discovery of objects, called *outliers*, that do not seem to have the characteristics of the general population. To quote Johnson [1]: "*an outlier is an observation in a data set which appears to be inconsistent with the remainder of that set of data*". For example, from a statistical point of view, an object is an outlier if it deviates significantly from the distribution.

Outlier discovery is performed for two main reasons: (i) removing the outliers before executing a clustering task leads to more effective cluster formation and (ii) outliers may not always be noise, but they may represent interesting elements that deserve further exploration (e.g., a large beautiful house sold in a very low price). Thus, either being noise or useful information, outliers should be mined efficiently.

One of the most widely used outlier definitions is the one based on distance: an object x is considered as an outlier, if there are less than k objects in a distance at most R from x, excluding x itself. On the other hand, if the number of objects

* Corresponding author.
E-mail addresses: kontaki@csd.auth.gr (M. Kontaki), gounaria@csd.auth.gr (A. Gounaris), papadopo@csd.auth.gr (A.N. Papadopoulos), tsichlas@csd.auth.gr (K. Tsichlas), manolopo@csd.auth.gr (Y. Manolopoulos).

http://dx.doi.org/10.1016/j.is.2015.07.006
0306-4379/

in the R-neighborhood of x is enough (i.e., more than k), then x is characterized as an *inlier*. The outliers defined this way are termed *distance-based* outliers [2,3], and the corresponding type of outlier detection has the advantages of detailed granularity of analysis and detecting isolated groups of outliers [6]. Note that, to characterize an object $x \in U$ as an outlier or an inlier, we just need a way to compute the distance between x and any other object of the universe U. If the objects are represented as points in a multi-dimensional space, then the distance can be any L_p norm (e.g., Euclidean or Manhattan). However, many applications require distance computations based on more expensive distance measures such as the Jaccard distance for near duplicate detection, the edit distance for sequence alignment in bioinformatics, distances based on quadratic form in multimedia applications and many more. Therefore, it is meaningful to provide general techniques that can work using many different distance measures and not to focus solely on multi-dimensional spaces.

Another important issue affecting the way outliers are mined is the dynamic nature of the universe U under consideration. In a static data set, we do not expect any changes in the outliers, since there are no insertions, deletions or updates. However, such data sets are rare in modern applications, which on the contrary require data mining tasks where changes are very frequent. Thus, in this work we focus on *data streams*, where the contents of U change continuously and, consequently, the set of outliers must be updated accordingly.

In data stream applications, data volumes are huge, meaning that it is not possible to keep all data memory resident. Instead, a *sliding window* is used, keeping a percentage of the data set in memory. The data objects maintained by the sliding window are termed *active objects*. When an object leaves the window we say that the object *expires*, and it is deleted from the set of active objects. There are two basic types of sliding windows: (i) the *count-based window* which always maintains the n most recent objects and (ii) the *time-based window* which maintains all objects arrived the last t time instances. In both cases, the *expiration time* of each seen object is known. The challenge is to design efficient algorithms for outlier monitoring, considering the expiration time of objects. Another important factor of stream-based algorithms is the memory space required for auxiliary information. Storage consumption must be kept low, enabling the possible enlargement of the sliding window, to accommodate more objects.

Contributions: In this work, we design efficient algorithms for continuous monitoring of distance-based outliers, in sliding windows over data streams, aiming at the elimination of the limitations of previously proposed algorithms. Our primary concerns are efficiency improvement and storage consumption reduction. A secondary concern stems from the problem of effective parameter configuration that application developers face; more specifically, it is hard to set R and k a priori in a way that meets the user needs. To address this, we allow for multiple configurations to be set and evaluated concurrently thus improving on the algorithm flexibility. In summary, our main contributions are the following:

- We prove a linear space lower bound which implies that in order to answer outlier queries on a set of objects one needs to store information about all objects even if we are settled with an approximate answer with a probability of success. This means that the window size W fully determines the number of stored objects, which are $O(W)$. This is a serious setback since in the various streaming models (e.g., [4]) we always strive for efficiency in queries as well as (asymptotic) minimization of space in order to support queries on larger sets of data. This is because the size W of the sliding window silently determines the size of the memory as well as the "interesting objects" to consider.
- A novel continuous algorithm is designed, which has two versions, and requires the radius R to be fixed but can handle multiple values of k. This algorithm (COD) consumes significantly less storage than previously proposed techniques and in addition, is more efficient.
- Since different users may have different views of outliers, we propose a new algorithm (ACOD) able to handle multiple values of k and multiple values of R, enabling the concurrent execution of different monitoring strategies.
- We propose an algorithm (MCOD) based on micro-clusters [5], to reduce the number of distance computations. There are cases where the distance function used is very expensive, and therefore, there is a need to keep this number low. This algorithm is also extended to support multiple queries (AMCOD).
- Performance evaluation results are offered based on real-life as well as synthetically generated data sets. The results show that our algorithms are consistently more efficient.

Roadmap: The rest of the work is organized as follows. Section 2 discusses related work in the area, whereas Section 3 presents some important preliminary concepts, to keep the article self-contained. In Section 4 we prove a lower bound on the space required to solve the problem. This bound essentially says that in order to monitor outliers in one pass we need to store information about all objects. We present our techniques in Section 5, whereas Section 6 contains the performance evaluation results based on real-life and synthetic data sets. Finally, Section 7 concludes the work and briefly discusses future work in the area.

2. Related work

Outlier detection is a topic that has been attracting the interest of researchers for several decades. Comprehensive surveys can be found in [6–10]. We distinguish between two main categories of techniques, static and steaming ones, which are discussed in turn.

2.1. Static outlier detection

Most of the early techniques originate from the statistics community [1,11], where the objects are modeled as a distribution, and objects are marked as outliers depending on their deviation from this distribution. However, for large dimensionalities and complex data types, statistical