MOOD: Moving Objects Outlier Detection

Salman Ahmed Shaikh and Hiroyuki Kitagawa

Graduate School of Systems and Information Engineering, University of Tsukuba
Tennodai, Tsukuba, Ibaraki 305-8573, Japan
salman@kde.cs.tsukuba.ac.jp, kitagawa@cs.tsukuba.ac.jp
http://www.kde.cs.tsukuba.ac.jp/

Abstract. This paper describes and demonstrates MOOD, a system for detecting outliers from moving objects data. In particular, we demonstrate a continuous distance-based outlier detection approach for moving objects' data streams. We assume that the moving objects are uncertain, as the state of a moving object can not be known precisely, and this uncertainty is given by the Gaussian distribution. The MOOD system provides an interface which takes moving objects' states streams and some parameters as input and continuously produces the distance-based outliers along with some graphs comparing the efficiency and accuracy of the underlying algorithms.

Keywords: Outlier Detection, Moving Objects, Uncertain Data, Data Streams.

1 Introduction

Outlier detection is a fundamental problem in data mining. It has applications in many domains including credit card fraud detection, network intrusion detection, environment monitoring, medical sciences, moving objects monitoring etc. Several definitions of outlier have been given in past, but there exists no universally agreed definition. Hawkins [1] defined an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Recently, with the advancement in data collection technologies, e.g., wireless sensor networks (WSN), data arrive continuously and contain certain degree of inherent uncertainty [2]. The causes of uncertainty may include but are not limited to limitation of equipments, inconsistent supply voltage and delay or loss of data in transfer [2]. Detection of outliers from such data is a challenging research problem in data mining. Hence this paper describes and demonstrates the MOOD system, to detect outliers from moving objects' streams, where the moving objects are uncertain and this uncertainty is given by the Gaussian distribution. In addition, the MOOD system generates dynamic graphs comparing the efficiency and accuracy of the underlying algorithms.

The problem of outlier detection on uncertain datasets was first studied by Aggarwal et al. [3]. However, their work was given for static data and cannot handle moving objects data or data streams. In [4], Wang et al. proposed an outlier

L. Chen et al. (Eds.): APWeb 2014, LNCS 8709, pp. 666–669, 2014.

Table 1. State sets

Time	State set	o_1	o_2	...	o_N
t_1	S^1	$\overrightarrow{\mathcal{A}_1^1}$	$\overrightarrow{\mathcal{A}_2^1}$	...	$\overrightarrow{\mathcal{A}_N^1}$
t_2	S^2	$\overrightarrow{\mathcal{A}_1^2}$	$\overrightarrow{\mathcal{A}_2^2}$	...	$\overrightarrow{\mathcal{A}_N^2}$
t_3	S^3	$\overrightarrow{\mathcal{A}_1^3}$	$\overrightarrow{\mathcal{A}_2^3}$	...	$\overrightarrow{\mathcal{A}_N^3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

detection approach for probabilistic data streams. However, their work focuses on tuple-level uncertainty. In contrast, in this work, attribute level uncertainty is considered.

2 Moving Objects Outlier Detection

This section describes the distance-based outlier detection approach for moving objects' data steams. In this paper, o_i denotes a k-dimensional uncertain object with attributes vector $\overrightarrow{\mathcal{A}_i} = (x_{i1}, ..., x_{ik})$ following the Gaussian distribution with mean $\overrightarrow{\mu_i} = (\mu_{i1}, ..., \mu_{ik})$ and co-variance matrix $\Sigma_i = diag(\sigma_{i1}^2, ..., \sigma_{ik}^2)$. Namely, $\overrightarrow{\mathcal{A}_i}$ is a random variable that follows the Gaussian distribution $\overrightarrow{\mathcal{A}_i} \sim \mathcal{N}(\overrightarrow{\mu_i}, \Sigma_i)$. Assuming that there are N objects whose states may change over time, $S^j = \{\overrightarrow{\mathcal{A}_1^j}, ..., \overrightarrow{\mathcal{A}_N^j}\}$ denotes a state set of N objects at time t_j. Note that the $\overrightarrow{\mu_i^j}$ denotes the observed coordinates (attribute values) of an object o_i at time t_j. Hence streams are the sequences of moving objects' states (locations) generated over time. We assume that the states of all the objects are generated synchronously and the set of states at a timestamp t_j is called a state set S^j as shown in table 1. We now present the definition of distance-based outliers for uncertain data streams, originally proposed in our previous work [6], as follows.

Definition 1. *An uncertain object o_i is a distance-based outlier at time t_j, if the expected number of objects in S^j lying within D-distance of o_i are less than or equal to threshold $\theta = N(1-p)$, where p is the fraction of objects that lie farther than D-distance of $o_i \in S^j$.*

The straightforward approach to detect outliers from each state set is to use the cell-based approach of uncertain distance-based outlier detection (UDB) for every timestamp, presented in our previous work [5]. However, the duration between two consecutive timestamps is usually very short and the state of all the objects may not change much in this duration. Hence, we proposed an incremental approach of outlier detection (denoted by CUDB in this work), which makes use of outlier detection results obtained from previous state set S^{j-1} at timestamp t_{j-1} to detect outliers in current state set S^j at timestamp t_j [6]. This eliminates the need to process all the objects' states at every timestamp and saves a lot of computation time.

3 The MOOD System Overview

In the following, we present an overview of the MOOD system, its interface and functionalities and the underlying algorithms.

3.1 The MOOD System Interface

The main interface of the MOOD system is shown in Fig. 1. The MOOD system accepts moving objects' streams. In addition, the system takes as input some parameters, as shown in the top-left corner of the MOOD system interface in Fig. 1. These parameters are required by the underlying algorithms for the computation of continuous outliers.

The MOOD system computes outliers for three algorithms. Namely UDB, CUDB and Knorr. In the MOOD system interface, the circles represent the objects at current timestamp while the squares represent the objects in the previous timestamp. The black bordered circles/squares represent the inliers. The red circles/squares represent the outliers identified by both the algorithms (i.e., CUDB and Knorr), while the pink and blue circles/squares represent the outliers identified by only CUDB algorithm and only Knorr algorithm, respectively. Moreover, the movement of objects is represented by green arrows as can be seen from Fig. 1.

3.2 The Underlying Algorithms

The UDB algorithm is proposed in our previous work [5] and can only detects distance-based outliers from uncertain static data. In order to make this algorithm work for continuous moving objects' data or streaming data, the UDB algorithm is executed for every timestamp. The CUDB is an incremental algorithm for distance-based outlier detection from uncertain data streams, proposed in our work [6]. The CUDB makes use of outlier detection results obtained from previous state set S^{j-1} at timestamp t_{j-1} to detect outliers from current state set S^j at timestamp t_j. This eliminates the need to process all the objects' states at every timestamp and saves a lot of computation time. The graph on the top-right corner of the MOOD interface (see Fig. 1) shows the difference in execution times of the UDB and the CUDB algorithms for moving objects data streams.

The Knorr algorithm is proposed by E.M.Knorr et al. [7] for distance-based outlier detection from deterministic static data. In the MOOD system, the Knorr outlier detection algorithm is used as a baseline to compare the accuracy of the CUDB algorithm. The graphs on the mid-right and the bottom-right of the MOOD interface (see Fig. 1) shows the comparison of precision and recall respectively. The precision is defined as the ability of the algorithms to present only true outliers. The recall is defined as the ability of the algorithms to present all true outliers. The precision and recall of the algorithms are measured on the perturbed dataset in the MOOD system. The perturbations are added to simulate the noise in moving objects' data. The perturbed dataset is obtained by adding normal random numbers with zero mean and standard deviation 15 to each of the tuple values of the original dataset.

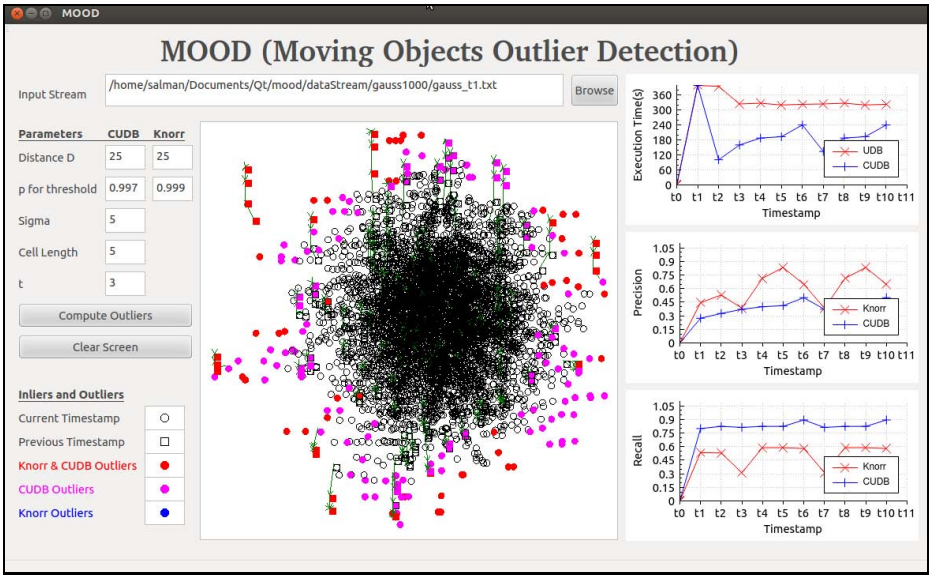

Fig. 1. The MOOD System Interface

Acknowledgements. This research was partly supported by the program "Research and Development on Real World Big Data Integration and Analysis" of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
2. Sharma, A.B., Golubchik, L., Govindan, R.: Sensor faults: detection methods and prevalence in real-world datasets. ACM Trans. Sens. Netw. 6(3), 23:1–23:39 (2010)
3. Aggarwal, C.C., Yu, P.S.: Outlier detection with uncertain data. In: SIAM ICDM, pp. 483–493 (2008)
4. Wang, B., Xiao, G., Yu, H., Yang, X.: Distance-based outlier detection on uncertain data. In: IEEE 9th ICCIT, pp. 293–298 (2009)
5. Shaikh, S.A., Kitagawa, H.: Efficient Distance-based Outlier Detection on Uncertain Datasets of Gaussian Distribution. In: World Wide Web, pp. 1–28 (2013)
6. Shaikh, S.A., Kitagawa, H.: Continuous Outlier Detection on Uncertain Data Streams. In: Proc. of IEEE 9th ISSNIP (2014)
7. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-Based Outliers: Algorithms and Applications. The VLDB Journal (2000)