Density-Based Local Outlier Detection on Uncertain Data*

Keyan Cao[1], Lingxu Shi[3], Guoren Wang[1,2], Donghong Han[1], and Mei Bai[1]

[1] College of Information Science & Engineering, Northeastern University, China
[2] Key Laboratory of Medical Image Computing (NEU), Ministry of Education
[3] Logistic Engineering University of People's Liberation Army, China
caokeyan@gmail.com

Abstract. Outlier detection is one of the key problems in the data mining area which can reveal rare phenomena and behaviors. In this paper, we will examine the problem of density-based local outlier detection on uncertain data sets described by some discrete instances. We propose a new density-based local outlier concept based on uncertain data. In order to quickly detect outliers, an algorithm is proposed that does not require the unfolding of all possible worlds. The performance of our method is verified through a number of simulation experiments. The experimental results show that our method is an effective way to solve the problem of density-based local outlier detection on uncertain data.

1 Introduction

Uncertainty is inherent to many important applications, such as location-based services (LBS), sensor monitoring and radio-frequency identification (RFID) [1,9]. In these applications, outlier detection often is essential when analyzing uncertain data. In many real-world applications, determining whether the object is an outlier, not only its distance to neighbors is considered, but also the density of surrounding neighbors should be considered. As such density-based outlier detection can consider local information [4].

Uncertain objects as referred to in this paper, by default, is in a d-dimensional numerical tuple. Given two uncertain instances $\widetilde{u_i}$ and $\widetilde{u_j}$, the distance between these is denoted by $d(\widetilde{u_i}, \widetilde{u_j})$. In this paper, we employ the Euclidian distance metric, but the developed techniques can be easily extended to other distance metrics.

Uncertain Objects. The uncertain object set U consists of $\{u_1, u_2,\cdots,u_i, \cdots, u_\mu\}$. Each uncertain object u_i has d dimensions. In this paper, we focus on the discrete case, an uncertain object u_i consists of a set of m instances $\{u_i^1, u_i^2, \cdots, u_i^j, \cdots, u_i^m\}$ where $1 \leq j \leq m$, $p(u_i^j)$ $(0 \leq p(u_i^j) \leq 1)$ denotes the probability of instance u_i^j occurring.

* This research are supported by the NSFC (Grant No. 61025007, 61328202, 61173029, 61100024, 61332006, 61073063), National High Technology Research and Development 863 Program of China (GrantNo.2012AA011004), National Basic Research Program of China (973, Grant No. 2011CB302200-G).

F. Li et al. (Eds.): WAIM 2014, LNCS 8485, pp. 67–71, 2014.

Definition 1. ***Distance sum of k-neighbors:*** *We assume that instance $\widetilde{u_j}$ is k-neighbor of instance $\widetilde{u_i}$ in a possible world W, and $n_k(\widetilde{u_i})$ is a k-neighbor set of $\widetilde{u_i}$ in W. Let $dis_k(\widetilde{u_i})$ denote the distance sum of k-neighbors of instance $\widetilde{u_i}$ in possible world W, then*

$$dis_k(\widetilde{u_i}) = \sum_{\widetilde{u_j} \in n_k(\widetilde{u_i})} dis(\widetilde{u_j}, \widetilde{u_i})$$

Definition 2. ***Density of an instance:*** *Let instances $\widetilde{u_j}$ and $\widetilde{u_i}$ be entities from different uncertain objects, i.e. $j \neq i$. If $\widetilde{u_j}$ is a k-neighbor of $\widetilde{u_i}$ in possible world W, the density of instance $\widetilde{u_i}$ is defined as*

$$den(\widetilde{u_i}) = \frac{k}{\sum_{W \in \mathbb{W}} dis_k(\widetilde{u_i}) P(W)}$$

Definition 3. ***k-neighbor set of an instance:*** *Let $\widetilde{u_i}$ denote an instance of uncertain object u_i, $(u_i \in U)$. $n_k(\widetilde{u_i})$ denotes the k-neighbor set of instance $\widetilde{u_i}$ in possible world W. Let $N_k(\widetilde{u_i})$ denote the k-neighbor set of instance $\widetilde{u_i}$ in all possible worlds, then*

$$N_k(\widetilde{u_i}) = \bigcup_{W \in \mathbb{W}} n_k(\widetilde{u_i})$$

Definition 4. ***Local Outlier Factor of an instance:*** *Given that instance $\widetilde{u_j}$ is a k-neighbor of instance $\widetilde{u_i}$, $den(\widetilde{u_j})$ denotes the density of $\widetilde{u_j}$, $LOF(\widetilde{u_i})$ denotes the LOF of instance $\widetilde{u_i}$.*

$$LOF(\widetilde{u_i}) = \frac{\sum_{\widetilde{u_j} \in n_k(\widetilde{u_i})} den(\widetilde{u_j}) P(W)}{k \times den(\widetilde{u_i})}$$

Definition 5. ***Local Outlier Factor of an uncertain object:*** *Let $\widetilde{u_i}$ denote any one instance of object u_i, $P(\widetilde{u_i})$ denotes the probability of $\widetilde{u_i}$, and $LOF(\widetilde{u_i})$ denotes the LOF of $\widetilde{u_i}$. $LOF(u_i)$ is then defined as*

$$LOF(u_i) = \sum_{\widetilde{u_i} \in u_i} LOF(\widetilde{u_i}) P(\widetilde{u_i})$$

Definition 6. ***Density-based local outlier:*** *When the uncertain objects are sorted in descending order based on their Local Outlier Factor, then the top-n uncertain objects are the density-based local outliers of the uncertain data set.*

2 Algorithms

In this section, we present an algorithm for Density-based Local Outlier detection on Uncertain data ($UDLO$), which transforms the outlier definition into a probability problem. Density based outliers can be calculated based on the definition in a naive way by finding all k-neighbors in all possible worlds. This solution however is impractical as the number of possible worlds grows exponentially with the number of instances. To overcome this, we propose an exact algorithm to compute the density of an instance.

Theorem 1. *Let $N_k(\widetilde{u_i})$ denote the k-neighbor set of instance $\widetilde{u_i}$. If all instances of an uncertain object are all in the k-neighbor set of instance $\widetilde{u_i}$, then object is a complete k-neighbor object of instance $\widetilde{u_i}$. The number of complete k-neighbor objects of instance $\widetilde{u_i}$ equals k. The maximal distance from instance to its complete k-neighbor object is denoted by $dis_c(\widetilde{u_i})$. The k-neighbor set of instance $\widetilde{u_i}$ consists of the instances whose distance to instance $\widetilde{u_i}$ is not larger than $dis_c(\widetilde{u_i})$.*

Theorem 2. *In the basic case, for $1 \leq i\ ,j \leq |N_k(\widetilde{u_j})|$, the order of instance $\widetilde{u_j}$ is t_i in list L. $P(S_{t_i}, \kappa)$ denotes the probability that there are κ instances existing in S_{t_i}. $P(S_{t_i}, 0) = P(S_{t_{i-1}}, 0)(1 - P(t_i)) = \prod_{j=1}^{i}(1 - P(t_i))$, and $P(S_{t_i}, \kappa) = P(S_{t_{i-1}}, \kappa - 1)P(t_i) + P(S_{t_{i-1}}, \kappa)(1 - P(t_i))$*

Theorem 3. *We assume that the instances $\widetilde{u_j}$ and $\widetilde{u_i}$ are from different uncertain objects, $\widetilde{u_j}$ is a k-neighbor of $\widetilde{u_i}$, and $P_k(\widetilde{u_j})$ denotes the probability, then the density of instance $\widetilde{u_i}$ is calculated as follows:*

$$den(\widetilde{u_i}) = \frac{k}{\sum_{\widetilde{u_j} \in N_k(\widetilde{u_i})} dis(\widetilde{u_j}, \widetilde{u_i}) P_k(\widetilde{u_j})}$$

Theorem 4. *If we assume that instance $\widetilde{u_j}$ is a k-neighbor of $\widetilde{u_i}$, then $P_k(\widetilde{u_j})$ is its k-neighbor probability. $den(\widetilde{u_i})$ and $den(\widetilde{u_j})$ are the densities of $\widetilde{u_i}$ and $\widetilde{u_j}$, then the LOF is given by:*

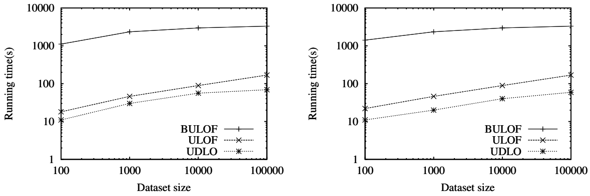
$$LOF(\widetilde{u_i}) = \frac{\sum_{\widetilde{u_j} \in N_k(\widetilde{u_i})} den(\widetilde{u_j}) P_k(\widetilde{u_j})}{k \times den(\widetilde{u_i})}$$

3 Experiments

We conducted several experiments on two real data sets and a synthetic data set to examine the efficiency and accuracy. In the remainder of this paper, these algorithms will be referred to as follows: outlier detection algorithm (denoted by $UDLO$). For comparison, we implemented the outlier detection algorithms (denoted by $BULOF$ and $ULOF$) which are proposed in [7].

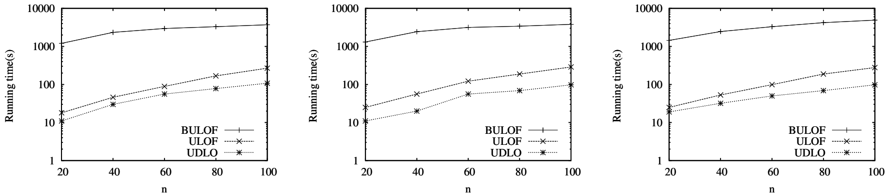
3.1 Efficiency

Efficiency is an important term frequently used in outlier detection studies. Figure 1 shows the results on the two different datasets. As expected $UDLO$ performs better than $ULOF$. Parameter n varies from 20 to 100, the running time increases as n increases. The running time of $ULOF$ is much higher than $UDLO$ algorithm.



(a) Synthetic (b) LDPA

Fig. 1. Running time vs. Data size



(a) Synthetic (b) LDPA (c) MAGIC

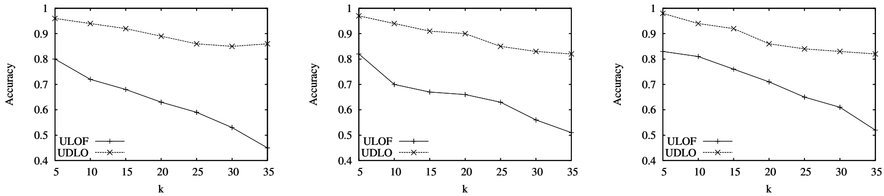
Fig. 2. Running time vs. n

3.2 Accuracy

In this section, we give the experimental results on accuracy, as shown in Figure 3. Since of the outliers lie in the center of a cluster, it is hard for the $ULOF$ algorithm to pick out this kind of objects from the entire dataset. The $UDLO$ algorithm adheres more strictly to the outlier definition, and therefore the accuracy of $UDLO$ algorithm is higher than that of $ULOF$. As expected, $UDLO$ can deliver the best results on all datasets.

4 Related Work

Aggarwal, C.C. *et al.* [2] were the first to investigate the problem of outlier detection on uncertain data. Wang *et al.*[8] focused on distance-based outlier detection on uncertain data, in which each data is affiliated with a confidence value. Jiang *et al.* [6] started with a comprehensive model considering both uncertain objects and their instances. In Ref.[3] they attempted to find outliers by building a global classifier. However, it is difficult to build a clear boundary



(a) Synthetic (b) LDPA (c) MAGIC

Fig. 3. Accuracy vs. k

between normal data and abnormal data. Fan [5] proposed density-based top-k outlier detection algorithm on uncertain objects. In their work, due to the distance between two objects is approximation, the density of object can not be accurate, so that affect the detection results. In Liu *et al.* [7], the authors proposed a signed outlier detection algorithm based on local information (local density and local uncertainty level) on uncertain data.

5 Conclusions

There are many important applications that require outlier detection on uncertain data. These applications always require that outlier are identified based on local information. We first derived an algorithm, which can effectively detect outliers without unfolding all possible worlds.

References

1. Aggarwal, C.: On density based transforms for uncertain data mining. In: ICDE, pp. 866–875 (2007)
2. Aggarwal, C., Yu, P.: Outlier detection with uncertain data. In: SDM, pp. 483–493 (2008)
3. Bo, L., Jie, Y., Shan, X.Y., Longbing, C., Philip, Y.: Exploiting local data uncertainty to boost global outlier detection. In: ICDM, pp. 304–303 (2010)
4. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: Lof:identifying density-based local outliers. Sigmod 29(2), 93–104 (2000)
5. Gaofeng, F., Hongmei, C., Zhiping, O.Y., Lizhen, W.: Density-based top-k outlier detection on uncertain objects. In: ICCSNT, vol. 4, pp. 2469–2472 (2011)
6. Jiang, B., Pei, J.: Outlier detection on uncertain data: objects, instances, and inferences. In: ICDE, pp. 422–433 (2011)
7. Liu, J., Deng, H.: Outlier detection on uncertain data based on local information. KBS (2013)
8. Wang, B., Xiao, G., Yu, H., Yang, X.: Distance-based outlier detection on uncertain data. CIT 1, 293–298 (2009)
9. Zhan, L., Zhang, Y., Zhang, W., Lin, X.: Finding top-k most influential spatial facilities over uncertain objects. In: CIKM, pp. 922–931 (2012)