Knowledge-Based Systems 51 (2013) 60–71

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Outlier detection on uncertain data based on local information

Jing Liu *, HuiFang Deng

Department of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, PR China

ARTICLE INFO

Article history:
Received 9 January 2013
Received in revised form 28 May 2013
Accepted 11 July 2013
Available online 20 July 2013

Keywords:
Outlier detection
Uncertain data
Local information

ABSTRACT

Based on local information: local density and local uncertainty level, a new outlier detection algorithm is designed in this paper to calculate uncertain local outlier factor (*ULOF*) for each point in an uncertain dataset. In this algorithm, all concepts, definitions and formulations for conventional local outlier detection approach (*LOF*) are generalized to include uncertainty information. The least squares algorithm on multi-times curve fitting is used to generate an approximate probability density function of distance between two points. An iteration algorithm is proposed to evaluate *K–η–distance* and a pruning strategy is adopted to reduce the size of candidate set of nearest-neighbors. The comparison between *ULOF* algorithm and the state-of-the-art approaches has been made. Results of several experiments on synthetic and real data sets demonstrate the effectiveness of the proposed approach.

1. Introduction

In recent years, a large amount of uncertain data are generated and collected due to new techniques of data acquisition, which are widely used in many real-world applications, such as wireless sensor networks, meteorology and mobile telecommunication. However, due to randomness and complexity caused by uncertain data, it is difficult to deal with using traditional data mining algorithms for deterministic data. Therefore, many researchers put efforts in developing new techniques of data processing and mining on uncertain data [2,6,24].

Outlier detection is one of the key problems in data mining area which can reveal rare phenomenon and behaviors, find interesting patterns, and have significant applied merits in many fields, such as detection of credit card fraud, network intrusion and abnormal climate. There are many published works on outlier detection [9,19,37]. Inclusion of uncertainty to the data makes the problem far more difficult to tackle, as this will further limits the accuracy of subsequent outlier detection. Therefore, how to effectively detect outliers on uncertain data is of great importance.

There are many challenges raised ahead to affect outlier detection on uncertain data:

How to fully employ uncertainty information of uncertain data? The uncertainty level of a point in an uncertain dataset refers to the degree to which the point is uncertain. The uncertainty level of an uncertain dataset refers to the degree to which the dataset is uncertain. Without considering uncertain information, even for an effective conventional method of outlier detection, it is hard to get accurate results on uncertain data. Let us use a 2-dimensional uncertain data set to explain. In Fig. 1, there are 12 circles of different radius and each circle indicates the range of possible value of a point. The larger the size of the circle, the higher the uncertainty of the point. If we apply the traditional local outlier detection algorithm (*LOF*) [8] to detect outlier on this dataset using distance expectation value to represent the distance between two different points, it is easy to recognize that the 8th-point is an outlier, because it is isolated from the neighbors, but the 9th-point would be probably missed as an outlier, as on the whole, the distance expectation value of it to its neighbors is roughly the same as normal points. However, its uncertainty level is relatively higher because the radius is larger than that of neighbors. As a result, there is a great probability that the 9th-point is an outlier. These observations show that the local uncertain behavior plays an important role in the process of detecting outlier on uncertain data. As it is stated in [5]. “In general, a higher level of uncertainty of a given record can result in it behaving like an outlier, even though the record itself may not be an outlier”. So it is meaningful to develop a new method which can fully employ local uncertainty behavior of uncertain data to detect outlier.

How to efficiently evaluate K **nearest-neighbor (*KNN*) for a given query point in uncertain data?** For deterministic data, the set of K nearest-neighbors for a given query point is clear and no ambiguity. But the situation on uncertain data is quite different. For example, we need to find 5-nearest-neighbors for query point q in Fig. 1. There are only three points (o_1,o_2,o_3) that are entirely contained within radius R while four other points (o_4,o_5,o_7,o_{10}) are partially covered. Obviously, points of o_1, o_2, o_3 are what we are searching for. But which would be the next two neighbors? In fact, any two points among the four points (o_4,o_5,o_7,o_{10}) have

* Corresponding author. Tel.: +86 20 87114028; fax: +86 20 87114638.
E-mail addresses: liujing.dm@qq.com (J. Liu), hdeng2008@gmail.com (H. Deng).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.07.005

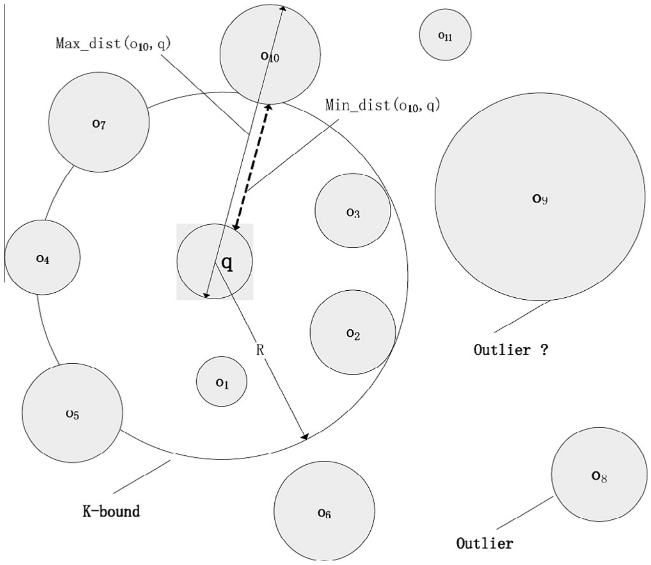

Fig. 1. The distribution of points in uncertain data.

a certain possibility to be the choice. Unfortunately, it needs an exponential amount of time to list every possible combination. This poses particular challenges for *KNN*-based algorithms, such as local outlier factor (*LOF*) algorithm. Therefore, how to effectively solve the problem of *KNN* query on uncertain data can be regarded as one of the most challenging problems in outlier detection on uncertain data.

Our contribution First, an outlier detection algorithm based on local information, termed as *ULOF* is proposed to assign an uncertain local outlier factor for each point in uncertain data set. According to the probability model of uncertain data, extended definitions which can exploit full uncertainty information are given in *ULOF* algorithm. Second, two optimization strategies are proposed to improve the efficiency of *K–η–distance* calculation and *K*-nearest neighbors selection. Third, characteristics of *ULOF* algorithm are analyzed in detail, and the influence of several mainly considered parameters on detection accuracy is investigated. At last, several experiments have been done to demonstrate the effectiveness and efficiency of *ULOF* algorithm.

The rest of this paper is organized as follows. In Section 2 we discuss related previous work. In Section 3 we give a detailed description of *ULOF* algorithm as well as two optimization strategies. The experimental results are presented in Section 4. The paper is concluded in Section 5.

2. Related work

There are many data mining techniques that have been proposed to solve corresponding problems on uncertain data, such as clustering [4,27], classification [30], frequent pattern mining [3] and probabilistic skylines [1,14,18]. In this section, we start with an introduction to two important models of uncertain data, and then give a brief review of previous works on outlier detection on deterministic data and uncertain data.

2.1. Uncertain data models

Models which are used to describe uncertain data are the basis for further research. Several approaches have been proposed to represent uncertainties. Fuzzy set theory [22], is an extension of classical set theory where set membership is defined as a possibility distribution. In possibility theory [23], uncertainty is described both by dual possibility and necessity measures. Rule-based modeling [7] is an approach that uses a set of rules to infer uncertainty and imprecision. The Dempster–Shafer theory of evidence [11], also known as the theory of belief functions, is an approach to obtain degrees of belief for one question from subjective probabilities for a related question and combine degrees of belief derived from independent items of evidence. Probabilistic models are widely used in many real applications to describe uncertain dataset. In this paper, we mainly focus on uncertainty in a probabilistic setting. Based on semantics of possible world [2,15], there are two main types of models of uncertain data: (1) attribute-level uncertainty model and (2) tuple-level uncertainty model.

In attribute-level uncertainty model, each attribute of a record is associated with an independent probability distribution function. In other words, the possible values for each attribute of a record are determined by a function that takes values based on a probability distribution. For example, when detecting air content, humidity and temperature information in a region, each of them would be subject to a certain independent probability distribution.

In tuple-level uncertainty model, each tuple is subject to its own probability of appearing in a possible world. These tuples are generally independent of each other as knowing one tuple would not be associated with others.

The attribute-level uncertainty model and tuple-level uncertainty model are equivalent in discrete case and can be transformed to each other [13,15]. Without loss of generality, we use attribute-level uncertainty model to describe uncertain data in this paper.

2.2. Outlier detection on uncertain data

There are many outlier detection methods on deterministic data in the literature. These techniques can be classified into five categories. They are statistical methods, distance-based methods, density-based methods, depth-based methods and angle-based methods. Statistical methods[12,25] (also known as model-based methods) assume that most data points are generated by a statistical model, while outliers do not fit the model. A distance-based outlier could be detected if there are not enough neighbor points within a specified distance [35,36]. A point is identified as density-based outlier if its density is relatively lower than that of surrounding points [8,20,28]. Depth-based approach [10,16] is based on the idea that outliers are located at the border of data region while normal points lie in the center. Angle-based methods [21,29] are especially useful for high-dimensional data, which argues that the variance of the angles between an outlier and other points is smaller than that of normal points.

Below, we will give a brief introduction to several studies in recent years of outlier detection on uncertain data:

Aggarwal et al. [5] first proposed a density-based outlier detection technique on uncertain data. They assumed that an outlier shows up in at least one subspace such that whose density is abnormally low. Therefore, to determine whether the point is an outlier, the density is calculated for each subspace where the point lies in. Formally, a point X is defined as a (δ,η)-*outlier* if it is lying in a subspace such that the probability of whose density being greater than η is lower than δ.

Wang et al. [34] first proposed distance-based outlier detection method on uncertain data. The tuple-level uncertainty model is used to describe uncertain dataset. Each tuple in uncertain dataset is affiliated with a confidence value to describe the probability of the tuple appearing in a possible world. A point in uncertain dataset can be considered as a distance-based outlier if the probability that we can find enough neighbor points within a specified distance is very low. The formal definition is that a point o is an $(up,$-