Neurocomputing 191 (2016) 224–237

Contents lists available at ScienceDirect

Neurocomputing

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

Probabilistic skyline queries on uncertain time series

Guoliang He [a,*], Lu Chen [a], Chen Zeng [a], Qiaoxian Zheng [b], Guofu Zhou [a]

[a] State Key Lab of Software Engineering, College of Computer Science, Wuhan University, Wuhan, China
[b] School of Computer Science and Information Engineering, Hubei University, Wuhan, China

ARTICLE INFO

Article history:
Received 15 November 2014
Received in revised form
6 December 2015
Accepted 6 December 2015
Communicated by P. Zhang
Available online 8 February 2016

Keywords:
Skyline query
Time series
Uncertainty

ABSTRACT

The uncertainty of data is popular and inherent in most applications. Although skyline queries on time series in the interval has attracted great interest in recent years, skyline queries on uncertain time series remains an open problem so far.

To handle this issue, we model the skyline queries on uncertain time series, and develop a two-step procedure to answer the probabilistic skyline queries on the dataset. First, three effective pruning techniques are proposed to obtain the skyline in the interval. Next, two simple methods are proposed to compute the skyline probability of each uncertain time series. For the online skyline queries, we also introduce a solution to improve the efficiency of pruning strategies by sharing the computation for two adjacent intervals. Experiments verify the effectiveness of probabilistic skylines and the efficiency and scalability of our algorithms.

1. Introduction

In the last decade there has been a dramatic explosion in the availability of measurements in a wide range of application domains, including traffic flow management [1], meteorology [2], astronomy [3], remote sensing [4,5], and object tracking [6]. Applications in the above domains usually organize these sequential measurements into time series, i.e., sequences of data points ordered along the temporal dimension, making time series a data type of particular importance. Meanwhile, the observed value at each timestamp exhibits various degree of uncertainty due to the limitations of measuring equipment, imcompleteness of data, environmental influence, etc. Due to the importance of these applications, analyzing uncertain time series has become an important task [7–9]. Particularly, some uncertain time series dominating others in a time interval are interesting and challenging.

Example 1. (motivation) An environmental investigation agency needs to analyze the Air Quality Index (AQI) of different regions in Beijing. Because the noise caused by the equipment itself and the influence of the surrounding environment always exist, sensor readings are inherently imprecise. To make the measured values approximating to the truth, several equipments are installed in each region. The readings of equipment over time can be captured by a time series. For illustration, here we assume each region has three equipments to monitor its AQI. Fig. 1 shows three synthesized uncertain time series of three regions in a monitoring period Jun 1–15, and each uncertain time series includes three instances. Suppose our task is to evaluate which region's environment is the worst in the interval from Jun 5 to Jun 11.

To handle this issue, an intuitive method is to compute the average readings of a region every day, as shown in Fig. 2. Then, traditional solving algorithms such as [10] could be adopted to obtain the most interesting region based on these certain time series, which consist of average readings every day. From Fig. 2 we could see that AQI of Region 2 is the worst since it has the highest average AQI in the query interval. However, using our proposed answering algorithm in this paper, the probability that AQI of each region is the worst are shown in Fig. 3.

As we know, the higher the probability of an uncertain time series example in the skyline, the worse the AQI of its Region. From Fig. 3 we can see that AQI of Region 1 is the worst because its skyline probability is the highest.

Due to the uncertainty of time series at each timestamp, a sample of an uncertain time series appears in a certain probability. To compare multiple uncertain time series in a query interval, it is interesting to find an uncertain time series and its probability that a sample of this uncertain time series is not dominated by any other uncertain time series. Technically, an uncertain time series s is in the probabilistic skyline on the query interval if there does not exist another uncertain time series s' such that any sample of s' is better than all possible samples of s on at least one timestamp and any sample of s' is not worse than all samples of s on other timestamps.

* Corresponding author.
E-mail address: glhe@whu.edu.cn (G. He).

http://dx.doi.org/10.1016/j.neucom.2015.12.104
0925-2312/

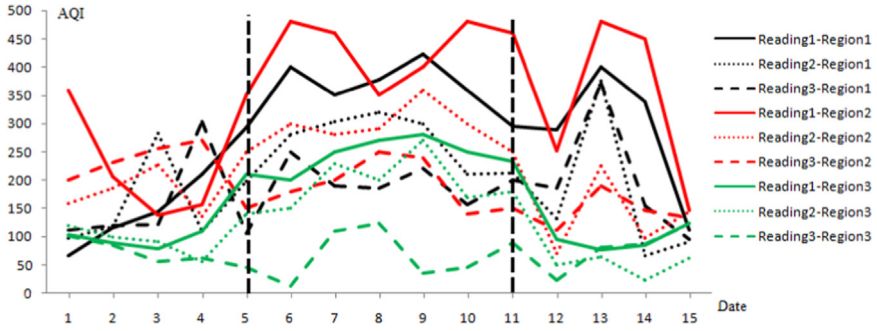

Fig. 1. A set of air quality index time series of three regions.

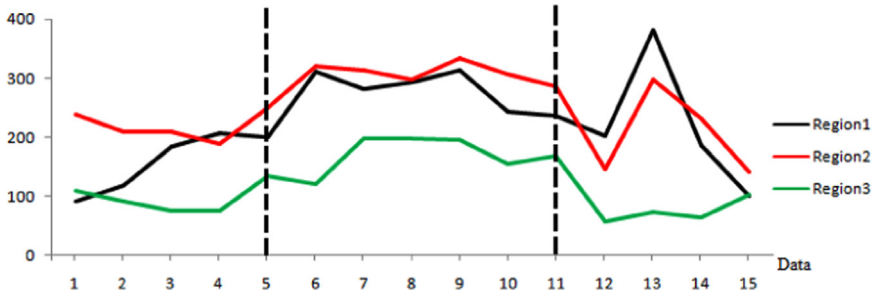

Fig. 2. The average air quality index time series of three regions.

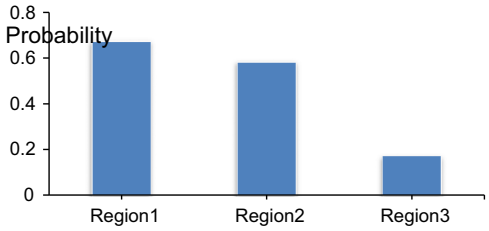

Fig. 3. The probabilities that AQI of each region is the worst in the interval.

We can easily give more applications where probabilistic skyline queries are useful. For instance, consider the rainfall of each month in different regions. Agricultural management institutions may be interested in some regions with high rainfall in a time interval. Due to the error or noise, Rainfall data is actually imprecise. Usually, several equipments are installed in a region. How to analyze these uncertain time series collected from these equipments to obtain the significant results is very critical in reality. Although some literatures have been researched the problem of skyline queries on time series [10,11], they assume that time series data is certain. Therefore, above methods are not suitable for the skyline queries on uncertain time series.

The notion of a probabilistic answer to skyline queries was first introduced in [12] for uncertain static data. After that many techniques and algorithms have been proposed to deal with the uncertain static data [13–15] and uncertain data streams [16–18]. However, uncertain time series is different from uncertain data stream. Generally, an uncertain data stream S is essentially a sequence of uncertain data sets continually produced over a time interval. Every instance is given a discrete time stamp, starting at 1 and ending at L, and write $S[t_1, t_2]$ for the uncertain data set consisting of all instances over a window covering time stamps t_1 to t_2. On the contrary, a time series U is defined as an ordered sequence over a time interval, and the element of each timestamp denotes the value of each attribute of this example. An uncertain time series means the element of each timestamp is a random variable modeling the real valued number, and it may have different values with a certain probability. Therefore, existing techniques answering skyline queries on uncertain data stream are not able to efficiently support uncertain time series on a sliding time window.

To the best of our knowledge, no prior work studied probabilistic skyline queries over uncertain time series. This is the first study about skyline analysis on uncertain time series. To handle the problem, we need to answer two essential questions. First, how can we model the probabilistic skyline on uncertain time series? Second, what is an efficient method to compute probabilistic skylines on uncertain time series?

In this paper, we make several contributions in answering the above questions. First, we model the skyline queries on uncertain time series and define some notations. Second, based on three novel pruning techniques, we propose two methods to answer the probabilistic skyline queries. For the issue of online skyline queries, we also introduce some techniques to improve the efficiency of pruning strategies by sharing the computation for two adjacent intervals. By abandoning the use of thresholds and computing the exact skyline probabilities of all uncertain time series in the skyline, our method allows more flexibility for users to utilize the skyline results according to their own utilities.

The rest of the paper is organized as follows. We review the related work in Section 2. In Section 3, we propose notions of probabilistic skylines on uncertain time series. In Section 4, we develop effective pruning techniques and algorithms for probabilistic skyline computation. An on-line skyline queries algorithm is introduced in Section 5. A systematic performance study is reported in Section 6. We conclude the paper in Section 7.

2. Related work

Since Borzsonyi et al. introduced the concept of skylines into the database field in 2001 [19], skyline queries [20–23] have been attracted by lots of researchers and many algorithms have been advanced to answer skyline queries on traditional certain data [24–26].

Pei et al. [12] first tackled the problem of skyline queries on uncertain data, and advanced bottom-up and top-down