Proc. 21st Int. Conf. on Scientific and Statistical Database Management (SSDBM'09), New Orleans, LA, 2009.

Probabilistic Similarity Search for Uncertain Time Series

Johannes Aßfalg, Hans-Peter Kriegel, Peer Kröger, Matthias Renz

Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 Munich, Germany
WWW: http://www.dbs.ifi.lmu.de
Email: {assfalg,kriegel,kroegerp,renz}@dbs.ifi.lmu.de

Abstract. A probabilistic similarity query over uncertain data assigns to each uncertain database object o a probability indicating the likelihood that o meets the query predicate. In this paper, we formalize the notion of uncertain time series and introduce two novel and important types of probabilistic range queries over uncertain time series. Furthermore, we propose an original approximate representation of uncertain time series that can be used to efficiently support both new query types by upper and lower bounding the Euclidean distance.

1 Introduction

Similarity search in time series databases is an active area of research usually with a focus on certain data. No work has been done so far to support query processing on uncertain time series. Uncertainty is important in emerging applications dealing e.g. with moving objects or object identification as well as sensor network monitoring. In all these applications, the observed values at each time slot of a time series exhibit various degrees of uncertainty. Due to the uncertainty of the data objects, similarity queries are probabilistic rather than exact: we can only assign to each database object a probability that it meets the query predicate. As a consequence, there is a need to adapt storage models and indexing/search techniques to deal with uncertainty [1–4]. Furthermore several approaches for probabilistic query processing have been proposed recently including probabilistic range queries [5, 6], probabilistic kNN and top-k queries [7–9, 2, 10] and probabilistic ranking [10–14]. Applications where the analysis of time series has to cope with uncertainty are e.g. traffic measurements in road networks, location tracking of moving objects or measuring environmental parameters as temperature.

When looking at the above sketched applications, we can extract two types of uncertain time series model uncertainty using a sampling approach rather than probability density functions (pdfs). In the first two applications, the sample values of different time slots are uncorrelated, i.e. there is no relationship between a given sample observation at time slot i and another sample observation at time slot $(i + 1)$. On the other hand, in Application 2, each observed sample at time slot i is correlated to an observation at time slot $(i+1)$ and *vice versa*. Since both

types require different and complex solutions in order to support probabilistic similarity queries, we only focus on uncorrelated uncertain time series throughout the rest of the paper. As indicated above, we assume that uncertainty is modelled using sample observations rather than pdfs.

To the best of our knowledge, this is the first paper, that formalizes the problem of probabilistic queries on uncertain time series, focusing on two types of probabilistic range queries (cf. Sec. 2). Furthermore, this paper proposes a novel compact approximation of uncertain time series and shows how upper and lower bounding distance estimations for Euclidean distance can be derived from these representations (cf. Sec. 3). Third, it illustrates how these distance approximations can be used to implement a multi-step query processor answering probabilistic similarity queries on uncertain time series efficiently (cf. Sec. 3).

2 Probabilistic Queries Over Uncertain Time Series

Usually, time series are sequences of (certain) d-dimensional points. Uncertain time series are sequences of points having an uncertain position in the d-dimensional vector space. This uncertainty is represented by a set of sample observations at each time slot.

Definition 1 (Uncertain Time Series). *An* uncertain time series $\mathcal{X}$ *of length* n *consists of a sequence* $\langle X_1, \ldots, X_n \rangle$ *of* n *elements, where each element* X_t *contains a set of* s d*-dimensional points (sample observations), i.e.* $X_t = \{x_{t,1}, \ldots, x_{t,s}\}$ *with* $x_{t,i} \in \mathbb{R}^d$*. We call* s *the* sample size *of* $\mathcal{X}$*. The distribution of the points in* X_t *reflects the uncertainty of* $\mathcal{X}$ *at time slot* t*.*

We will use the term *regular time series* for traditional, non-uncertain (i.e. exact) time series consisting of only one d-dimensional point at each time slot [1].

In order to measure the similarity of uncertain time series we need a distance measure for such uncertain time series. For regular time series, e.g. any L_p-norm is commonly used to measure the distance between pairs of time series. Due to the uncertainty of the time series, also the distance between two time series is uncertain. Instead of computing one unique distance value such as the L_p-norm of the corresponding sequences, the distance between uncertain time series rather consists of multiple distance values reflecting the distribution of all possible distance values between the samples of the corresponding uncertain time series. This intuition is formalized in the following definition.

Definition 2 (Uncertain L_p-Distance). *For a one-dimensional uncertain time series* $\mathcal{X}$ *of length* n*, let* $s_\mathcal{X}$ *be the sample size of* $\mathcal{X}$ *and* $TS_\mathcal{X}$ *be the set of all possible regular time series that can be derived from the combination of different sample points of* $\mathcal{X}$ *by taking one sample from each time slot, i.e.*

$$TS_\mathcal{X} = \{\langle x_{1,1}, x_{2,1}, \ldots, x_{n,1} \rangle, \ldots, \langle x_{1,s_\mathcal{X}}, x_{2,s_\mathcal{X}}, \ldots, x_{n,s_\mathcal{X}} \rangle\}.$$

[1] For presentation issues, we assume 1-dimensional uncertain time series, the extension to the general d-dimensional case is straightforward.

The L_p-distance between two uncertain time series $\mathcal{X}$ *and* $\mathcal{Y}$*, denoted by* $\widetilde{dist}_p$*, is a collection containing the* L_p *distances of all possible combinations from* $TS_\mathcal{X}$ *and* $TS_\mathcal{Y}$*, i.e.* $\widetilde{dist}_p(\mathcal{X},\mathcal{Y}) = \{L_p(x,y) \,|\, x \in TS_\mathcal{X}, y \in TS_\mathcal{Y}\}$.

Based on the distance function $\widetilde{dist}_p$ we define two query types for uncertain time series. Thereby, we define the probability $\Pr(\widetilde{dist}_p(\mathcal{X},\mathcal{Y}) \leq \varepsilon)$ that the distance between two uncertain time series $\mathcal{X}$ and $\mathcal{Y}$ is below a given threshold ε as

$$\Pr(\widetilde{dist}_p(\mathcal{X},\mathcal{Y}) \leq \varepsilon) = \frac{|\{d \in \widetilde{dist}_p(\mathcal{X},\mathcal{Y}) | d \leq \varepsilon\}|}{s^n_\mathcal{X} \cdot s^n_\mathcal{Y}}.$$

Definition 3 (Probabilistic Bounded Range Query). *Let* $\mathcal{D}$ *be a database of uncertain time series,* $\varepsilon \in \mathbb{R}^+$*, and* $\tau \in [0,1]$*. For an uncertain time series* $\mathcal{Q}$*, the* Probabilistic Bounded Range Query *(PBRQ) returns the following set*

$$RQ_{\varepsilon,\tau}(\mathcal{Q},\mathcal{D}) = \{\mathcal{X} \in \mathcal{D} \,|\, \Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) \geq \tau\}.$$

Definition 4 (Probabilistic Ranked Range Query). *Let* $\mathcal{D}$ *be a database of uncertain time series and* $\varepsilon \in \mathbb{R}^+$*. For an uncertain query time series* $\mathcal{Q}$*, the* Probabilistic Ranked Range Query *(PRRQ) returns an ordered list:*

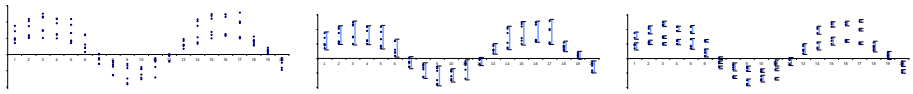
$$RQ_{\varepsilon,rank}(\mathcal{Q},\mathcal{D}) = (\mathcal{X}_1,\ldots,\mathcal{X}_m),$$

where $\Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}_i) \leq \Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}_{i+1})$ *(*$1 \leq i \leq, m-1$*) and* $\Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}_i) \leq \varepsilon)$ *for all* $i = 1,\ldots,m$*. For efficiency reasons, we assume a function getNext on the set* $RQ_{\varepsilon,rank}(\mathcal{Q},\mathcal{D})$ *that returns the next element of the ranking, i.e. the first call of getNext returns the first element in* $RQ_{\varepsilon,rank}(\mathcal{Q},\mathcal{D})$*, the second call returns the second element in* $RQ_{\varepsilon,rank}(\mathcal{Q},\mathcal{D})$*, and so on.*

Let us note that in the database context where we have long time series (high value of n) and high sample rates, the naive solution for both query types are CPU-bound because for all $\mathcal{X} \in \mathcal{D}$ we need to compute all distance observations in $\widetilde{dist}_p(\mathcal{Q},\mathcal{X})$ in order to determine $\Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$. This means that a naive solution requires to compute for each $\mathcal{X} \in \mathcal{D}$ exactly $|\widetilde{dist}_p(\mathcal{Q},\mathcal{X})| = s^n_\mathcal{Q} \cdot s^n_\mathcal{X}$ distances. For large values of n, $s_\mathcal{Q}$, and $s_\mathcal{X}$, this is obviously much more costly than sequentially scanning the disk to access all $\mathcal{X} \in \mathcal{D}$.

3 Multi-Step Probabilistic Range Query Processing

Obviously, the CPU cost (and thus, the overall runtime) of our probabilistic similarity queries are dominated by the number of distance calculations necessary to determine the probability $\Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$ for a query object $\mathcal{Q}$ and all $\mathcal{X} \in \mathcal{D}$. This high number results from the combination of the observed distance values between $\mathcal{Q}$ and $\mathcal{X}$ at each time slot. A first idea for runtime reduction is that we only need to determine the number of distance observations



(a) Exact level. (b) First level. (c) Second level.

Fig. 1. Different levels of approximating uncertain time series.

$d \in \widetilde{dist}_p(\mathcal{Q}, \mathcal{X})$ with $d \leq \varepsilon$ because $|\widetilde{dist}_p(\mathcal{Q}, \mathcal{X})| = s^n_{\mathcal{Q}} \cdot s^n_{\mathcal{X}}$. We can further improve the runtime by calculating lower and upper bounds for the probability that further reduce the number of distance computations. For that purpose, we have to calculate an upper and a lower bound for the number of distance observations $d \in \widetilde{dist}_p(\mathcal{Q}, \mathcal{X})$ with $d \leq \varepsilon$.

3.1 Approximative Representation

Intuitively, we construct the approximative representation of an uncertain time series $\mathcal{X}$ by aggregating the observations $x_{i,j} \in X_i$ at each time slot i into groups and use these groups to calculate the distance between uncertain time series. Obviously, this reduces the sample rate and thus, the overall number of possible distance combinations. The groups are represented by minimum bounding intervals[2].

Definition 5 (Approximative Representation). *The* approximative representation $\mathcal{X}_a$ *of an uncertain time series* $\mathcal{X}$ *of length* n *consists of a sequence* $\langle \{I_{1,1}, \ldots, I_{1,m_1}\}, \ldots, \{I_{n,1}, \ldots, I_{n,m_n}\} \rangle$ *of interval sets. Each interval* $I_{i,j} = [l_{i,j}, u_{i,j}]$ *minimally covers a given number* $|I_{i,j}|$ *of sample points of* X_i*, i.e.* $l_{i,j}$ *and* $u_{i,j}$ *are sample points of* X_i*, at time slot* i*.*

We use two levels of approximation. The first level describes all sample points at time slot i by one minimal bounding interval (cf. Figure 1(b)), i.e. $m_i = 1$ for all time slots i and $\mathcal{X}_a = \langle I_{1,1}, \ldots, I_{n,1} \rangle$. For the second level approximations, the sample observations at time slot i are grouped into k clusters by applying the algorithm k-means [15] on all $x_{i,j} \in X_i$ (cf. Figure 1(c)), i.e. $m_i = k$ for all time slots i and $\mathcal{X}_a = \langle \{I_{1,1}, \ldots, I_{1,k}\}, \ldots, \{I_{n,1}, \ldots, I_{n,k}\} \rangle$.

3.2 Distance Approximations

Using approximative representations $\mathcal{X}_a$ and $\mathcal{Y}_a$ of two uncertain time series $\mathcal{X}$ and $\mathcal{Y}$ we are able to calculate lower and upper bounds for $\Pr(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon)$.

Analogously to Definition 2, let $TS_{\mathcal{X}_a}$ be the set of all possible approximated regular time series derived from the combination of different intervals of $\mathcal{X}_a$ by

[2] or minimum bounding hyper-rectangles in the d-dimensional case

taking one interval from each time slot, i.e.

$$TS_{\mathcal{X}_a} = \{\langle I_{1,1}, I_{2,1}, \ldots, I_{n,1}\rangle, \ldots, \langle I_{1,l_1}, \ldots, I_{n,l_n}\rangle\}.$$

Let $X_a \in TS_{\mathcal{X}_a}$ and let $[l_{x_i}, u_{x_i}]$ be the interval of X_a at time slot i. The distance $L_{L_p}(X_a, Y_a) = \sqrt[p]{\sum_{i=1}^{n} (\max\{0, \max\{l_{x_i}, l_{y_i}\} - \min\{u_{x_i}, u_{y_i}\}\})^p}$ is the smallest L_p-distance between all intervals of $X_a \in TS_{\mathcal{X}_a}$ and $Y_a \in TS_{\mathcal{Y}_a}$, whereas the distance
$U_{L_p}(X_a, Y_a) = \sqrt[p]{\sum_{i=1}^{n} (\max\{u_{x_i} - l_{y_i}, u_{y_i} - l_{x_i}\})^p}$ is the largest L_p-distance between all intervals of $X_a \in TS_{\mathcal{X}_a}$ and $Y_a \in TS_{\mathcal{Y}_a}$. Aggregating these distance values by means of the distance function $\widetilde{dist}_p$, we obtain an interval of distances bound by L_{dist} and U_{dist}. Now, we can lower bound each distance observation in $\widetilde{dist}_p(\mathcal{X}, \mathcal{Y})$ by
$LB_p(\mathcal{X}_a, \mathcal{Y}_a) = \{(L_{dist}(X_a, Y_a))^{|X_a|\cdot|Y_a|} | X_a \in TS_{\mathcal{X}_a}, Y_a \in TS_{\mathcal{Y}_a}\}$.
Analogously, we can upper bound each distance observation in $\widetilde{dist}_p(\mathcal{X}, \mathcal{Y})$ by
$UB_p(\mathcal{X}_a, \mathcal{Y}_a) = \{(U_{dist}(X_a, Y_a))^{|X_a|\cdot|Y_a|} | X_a \in TS_{\mathcal{X}_a}, Y_a \in TS_{\mathcal{Y}_a}\}$.

Lemma 1. *Let $X_a = \langle I_1^x, \ldots, I_n^x\rangle \in TS_{\mathcal{X}_a}$ and $\mathcal{Y}_a = \langle I_1^y, \ldots, I_n^y\rangle \in TS_{\mathcal{Y}_a}$ be approximated regular time series. For all $x = \langle x_1, \ldots, x_n\rangle$, $x_i \in I_i^x$ and for all $y = \langle y_1, \ldots, y_n\rangle$, $y_i \in I_i^y$, the following inequalities hold:*

$$L_{L_p}(\mathcal{X}_a, \mathcal{Y}_a) \leq L_p(x, y).$$

$$U_{L_p}(\mathcal{X}_a, \mathcal{Y}_a) \geq L_p(x, y).$$

A lower bound of the probability $\Pr(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon)$ can be defined as

$$\Pr_{LB}(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon) = \frac{|\{d \in UB_p(\mathcal{X}_a, \mathcal{Y}_a) | d \leq \varepsilon\}|}{s_{\mathcal{X}}^n \cdot s_{\mathcal{Y}}^n}$$

and an upper bound as

$$\Pr_{UB}(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon) = \frac{|\{d \in LB_p(\mathcal{X}_a, \mathcal{Y}_a) | d \leq \varepsilon\}|}{s_{\mathcal{X}}^n \cdot s_{\mathcal{Y}}^n}$$

Lemma 2. *For any uncertain time series $\mathcal{X}$ and $\mathcal{Y}$, the following inequations hold:*

(1) $\Pr_{LB}(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon) \leq \Pr(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon)$

(2) $\Pr_{UB}(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon) \geq \Pr(\widetilde{dist}_p(\mathcal{X}, \mathcal{Y}) \leq \varepsilon)$

The proofs of Lemma 1 and 2 can be found in [16], but are omitted here due to space limitations.

The two following query types are based on an iterative filter-refinement policy. A queue Q_{Ref} is used to organize all uncertain time series sorted by descending upper bounding probabilities $\Pr_{UB}(\widetilde{dist}_p(\mathcal{Q}, \mathcal{X}) \leq \varepsilon)$ w.r.t. the query object $\mathcal{Q}$.

3.3 Probabilistic Bounded Range Queries (PBRQ)

In an iterative process we remove the first element $\mathcal{X}$ of the queue Q_{ref}, compute its lower and upper bounding probabilities $\Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$ and $\Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$, and filter $\mathcal{X}$ according to these bounds. If $\Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) \geq \tau$, then $\mathcal{X}$ is a true hit and is added to the result set. If $\Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) < \tau$, then $\mathcal{X}$ is a true drop and can be pruned. Otherwise, $\mathcal{X}$ has to be refined. Let us note that we do not immediately refine the object completely. Rather, the refinement is performed in several steps (1st level to 2nd level, 2nd level to exact representation). Details on the strategies for the step-wise refinement are presented below in Section 3.5. After the partial refinement step, $\mathcal{X}$ is again inserted into Q_{Ref} if it cannot be pruned or reported as true hit according to the above conditions and is not refined completely yet. If an object $\mathcal{X}$ is refined completely, then obviously $\Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) = \Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) = \Pr(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$. The iteration loop stops if Q_{Ref} is empty, i.e. all objects are pruned, identified as true hits before complete refinement, or are completely refined.

3.4 Probabilistic Ranking Range Query (PRRQ)

After initialization, the method *getNext()* can be called, returning the next object in the ranking. Obviously, an object $\mathcal{X}$ is the object with the highest probability if for all objects $\mathcal{Y} \in \mathcal{D}$ the following property holds: $\Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) \geq \Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{Y}) \leq \varepsilon)$. Since the candidate objects of the database are ordered by descending upper bounding probabilities in Q_{Rank}, we only need to check if the lower bounding probability of the first element in Q_{Rank} is greater or equal to the upper bounding probability of the second element. If this test returns true, we can report the first object as the next ranked object. Otherwise, we have to refine the first object in Q_{Rank} in order to obtain better probability bounds. As discussed above, this refinement is step-wise, i.e. several refinement steps are necessary in order to obtain the exact probability. The idea of the method *getNext()* is to iteratively refine the first object in Q_{Rank} as long as the lower bounding probability of this element is lower than the upper bounding probability of the second element in Q_{Rank}.

3.5 Step-Wise Refinement of Probability Estimations

The aim for each refinement step is to be able to identify an uncertain time series as true hit or true drop. This aim is reached for an uncertain time series $\mathcal{X}$ if the probability interval $[\Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon), \Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)]$ is above or below τ. For this reason, we try to increase the lower bound of the probability $\Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$ in the case that

$$\tau - \Pr_{LB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) \leq \Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon) - \tau$$

holds. Otherwise, we try to decrease $\Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$.

For increasing the lower bounds of the probabilities for $\mathcal{X}$ we have to refine those intervals which are intersected by the ε value such that we refine first that approximated distance which probably will be resolved into a set of approximated distances that are clearly below ε and approximates as many distances $d \in \widetilde{dist}_p(\mathcal{Q},\mathcal{X})$ as possible. Here we use the following heuristic: The increase of the number of detected distances $d \in \widetilde{dist}_p(\mathcal{Q},\mathcal{X})$ that are clearly below ε can be estimated by

$$\widetilde{w} = (1 - \frac{s_u}{\max_{i=1..n}\{d_{u,i} - d_{l,i}\}}) \cdot |X_a| \cdot |Q_a|,$$

where $s_u = U_{dist}(Q_a, X_a) - \varepsilon$, $d_{u,i} = \max\{u_{q_i} - l_{x_i}, u_{x_i} - l_{q_i}\}$, $d_{l,i} = \max\{0, \max\{l_{q_i}, l_{x_i}\} - \max\{u_{q_i}, u_{x_i}\}\}$ and $|X_a| \cdot |Q_a|$ corresponds to the number of distances which are approximated by $U_{dist}(Q_a, X_a)$ and $L_{dist}(Q_a, X_a)$. The example depicted in Figure 2 shows the situation of the approximated distance $\widetilde{d} = (L_{L_1}(Q_a, X_a), U_{L_1}(Q_a, X_a))$ before (top) and after (bottom) the refinement step. The approximated distance $\widetilde{d}$ is refined by refining exactly one of the n distance intervals in the time domain that correspond to $\widetilde{d}$. Obviously, the number of distances approximated by $\widetilde{d}$ is the product of the number $|Q_a|$ of regular time series approximated by Q_a and the number $|X_a|$ of regular time series approximated by X_a. In order to estimate the number of approximated distances that fall below ε after refining $\widetilde{d}$, we have to look at the distance intervals in the time domain. When refining a distance interval in the time domain, e.g. $(d_{l,5}, d_{u,5})$ in our example, then all resulting distance intervals that are clearly below $d_{u,i} - s_u$ correspond to the resulting approximated distances that are below ε. Since $\widetilde{w}$ has to be maximized, we should refine $\widetilde{d}$ by refining the largest time interval in the time domain. Finally, based on the described estimation, we refine the approximated distance for which $\widetilde{w}$ is maximal. In the case we want to decrease the upper bound of the probability $\Pr_{UB}(\widetilde{dist}_p(\mathcal{Q},\mathcal{X}) \leq \varepsilon)$ we can use a very similar refinement strategy.

4 Summary of Experimental Results

In this short proposal, we just want to give a brief summary of our experimental results due to limited space. For the interesting reader we refer to [16] where a broader discussion of our experiments can be found. Our datasets are based on several artificial and real-world benchmark datasets derived from a wide range, including *CBF*, *GunX*, *SynCtrl* and *Leaf* from the UCI Time Series Data Mining Archive[3]. The time series are modified to get uncertain time series by means of sampling around the given exact time series values according to specific distribution functions (e.g. uniform and Gaussian). As discussed above, the computation of probabilistic similarity queries is CPU-bounded. To achieve a fair comparison which is independent of the implementation, we measured the

[3] http://kdd.ics.uci.edu/

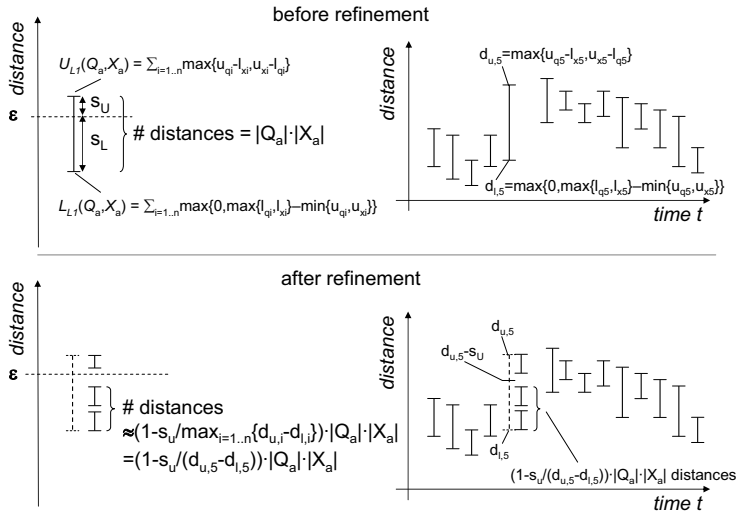


Fig. 2. Refinement heuristics.

efficiency by the average number of required calculations required to execute a query.

At first we measured the speed-up factor our approach yields compared to the straightforward approach naively computed as defined in Section 2. In the first experiment, we examine how our approach can speed up PBRQ and PRRQ for different datasets and varying sample rates. The speed-up factor of both query types is between 10^{75} and 10^{300} and increases exponentially with linearly increasing the sample rate. The rational for this is that the number of possible time-series instances increases exponentially with the time series length and the number of samples used for each time slot. Furthermore, we could show that our approach scales significantly better than the competitor w.r.t. the database size. Finally, we could experimentally show that our refinement strategy clearly outperforms more simple refinement strategies and that this superiority of our approach is robust w.r.t. all query parameters.

5 Conclusions

To the best of our knowledge, we propose the first approach for performing probabilistic similarity search on uncertain time series in this paper. In particular, we formalize the notion of uncertain time series and introduce two novel probabilistic query types for uncertain time series. Furthermore, we propose an original method for efficiently supporting these probabilistic queries using a filter-refinement query processing.

References

1. Benjelloun, O., Sarma, A.D., Halevy, A., Widom, J.: "ULDBs: Databases with Uncertainty and Lineage". In: Proc. 32th Int. Conf. on Very Large Data Bases (VLDB'06), Seoul, Korea. (2006) 1249–1264
2. Re, C., Dalvi, N., Suciu, D.: "Efficient top-k query evaluation on probalistic databases". In: Proc. 23th Int. Conf. on Data Engineering (ICDE'07), Istanbul, Turkey. (2007)
3. Sen, P., Deshpande, A.: "Representing and querying correlated tuples in probabilistic databases". In: Proc. 23th Int. Conf. on Data Engineering (ICDE'07), Istanbul, Turkey. (2007)
4. Antova, L., Jansen, T., Koch, C., Olteanu, D.: "Fast and Simple Relational Processing of Uncertain Data". In: Proc. 24th Int. Conf. on Data Engineering (ICDE'08), Cancún, México. (2008)
5. Cheng, R., Xia, Y., Prabhakar, S., Shah, R., Vitter, J.: "Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data". In: Proc. 30th Int. Conf. on Very Large Data Bases (VLDB'04), Toronto, Cananda. (2004) 876–887
6. Kriegel, H.P., Kunath, P., Pfeifle, M., Renz, M.: "Probabilistic Similarity Join on Uncertain Data". In: Proc. 11th Int. Conf. on Database Systems for Advanced Applications, Singapore, pp. 295-309. (2006)
7. Cheng, R., Kalashnikov, D., Prabhakar, S.: "Evaluating Probabilistic Queries over Imprecise Data". In: Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'03), San Diego, CA, USA). (2003) 551–562
8. Kriegel, H.P., Kunath, P., Renz, M.: "Probabilistic Nearest-Neighbor Query on Uncertain Objects". In: Proc. 12th Int. Conf. on Database Systems for Advanced Applications, Bangkok, Thailand. (2007)
9. Lian, X., Chen, L.: "Probabilistic ranked queries in uncertain databases". In: EDBT 2008, 11th International Conference on Extending Database Technology, Nantes, France, March 25-29, 2008, Proceedings. (2008) 511–522
10. Yi, K., Li, F., Kollios, G., Srivastava, D.: "Efficient Processing of Top-k Queries in Uncertain Databases with x-Relations". IEEE Trans. Knowl. Data Eng. **20**(12) (2008) 1669–1682
11. Böhm, C., Pryakhin, A., Schubert, M.: "Probabilistic Ranking Queries on Gaussians". In: Proc. 18th Int. Conf. on Scientific and Statistical Database Management (SSDBM'06), Vienna, Austria. (2006) 169–178
12. Cormode, G., Li, F., Yi, K.: "Semantics of Ranking Queries for Probabilistic Data and Expected Results". In: Proc. 25th Int. Conf. on Data Engineering (ICDE'09), Shanghai, China. (2009) 305–316
13. Tao, Y., Cheng, R., Xiao, X., Ngai, W., Kao, B., Prabhakar, S.: "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions". In: Proc. 31th Int. Conf. on Very Large Data Bases (VLDB'05), Trondheim, Norway. (2005) 922–933
14. Soliman, M., Ilyas, I.: "Ranking with Uncertain Scores". In: Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29-April 2, 2009, Shanghai, China. (2009) 317–328
15. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5-th Berkeley Symposium on Mathematical Statistics and Probability. (1967)
16. Assfalg, J., Kriegel, H.P., Kröger, P., Renz, M.: "Probabilistic Similarity Search for Uncertain Time Series". Tech. Rep. 2009: http://www.dbs.ifi.lmu.de/ renz/technicalReports/uncertainTimeSeries.pdf