Uncertainty in Sequential Pattern Mining

Muhammad Muzammal and Rajeev Raman

Department of Computer Science, University of Leicester, UK.
{mm386,r.raman}@mcs.le.ac.uk

Abstract. We study uncertainty models in *sequential pattern mining*. We discuss some kinds of uncertainties that could exist in data, and show how these uncertainties can be modelled using probabilistic databases. We then obtain *possible world semantics* for them and show how frequent sequences could be mined using the *probabilistic frequentness* measure.

Key words: Mining Uncertain Data, Sequential Pattern Mining, Probabilistic Databases, Theoretical Foundations of Data Mining.

1 Sequential Pattern Mining

Sequential pattern mining [2] is an important data mining problem: it is concerned with databases that contain sequences of *events*, each of which is associated with a *source*. For example, a transaction database of a store may contain sequences of purchases (events) made by individual customers (sources), and the objective is to find patterns of customer purchasing behaviour in successive visits. This has applications in various domains including transaction databases, web access patterns and biological sequences, and is formally defined as follows. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_q\}$ be a set of *items* and $\mathcal{S} = \{1, \ldots, m\}$ be a set of *sources*. An *event* $e \subseteq \mathcal{I}$ is a collection of items. A *database* $D = \langle r_1, r_2, \ldots, r_n \rangle$ is an ordered list of *records* such that each $r_i \in D$ is of the form (eid_i, e_i, σ_i), where eid_i is event-id, e_i is an event and σ_i is a source. A *sequence* $s = \langle s_1, s_2, \ldots, s_a \rangle$ is an ordered list of events. Let $s = \langle s_1, s_2, \ldots, s_q \rangle$ and $t = \langle t_1, t_2, \ldots, t_r \rangle$ be two sequences. We say that s is a *subsequence* of t, denoted $s \preceq t$, if there exist integers $1 \leq i_1 < i_2 < \cdots < i_q \leq r$ such that $s_k \subseteq t_{i_j}$, for $k = 1, \ldots, q$. The *source sequence* corresponding to a source i, denoted by D_i, is just the multiset $\{e|(eid, e, i) \in D\}$, ordered by *eid*. For a sequence s and source i, let $X_i(s, D)$ be an indicator variable, whose value is 1 if $s \preceq D_i$, and 0 otherwise. The objective is to find all sequences s whose *support* ($Supp$) is at least some user-defined threshold $\theta, 1 \leq \theta \leq m$, where $Supp(s, D) = \sum_{i=1}^{m} X_i(s, D)$.

2 Modelling Uncertainty

Traditionally, it is assumed that data is deterministic. However, it is now recognized that data is often inherently noisy or uncertain. *Probabilistic* databases are one way to model such uncertainties [1, 7]. Recently, many data mining problems

have been studied in probabilistic databases including frequent itemset mining [1, 3]. We focus on sequential pattern mining and our interest is in situations where there is uncertainty either about a source or in the associated events.

Source-Level Uncertainty. In a retail transaction database, a customer's details may be incomplete or incorrect, or the database may itself be uncertain as a result of "deduplication" or cleaning [4], leading to ambiguity in the customer's identity. A person/vehicle may be detected by a sensor/camera, but identification methods may be noisy, leading to uncertainty (take the UK police's automatic number plate recognition database [9] for example). In such scenarios, it is certain that an event occurred (e.g. a customer bought some items, a vehicle/person entered an area) but there is uncertainty about the source associated with that event. Situations like this can be modelled using *attribute level* uncertainty [7], when the 'source' attribute is a probability distribution over sources.

A *probabilistic database* D^p is an ordered list $\langle r_1, \ldots, r_n \rangle$ of records of the form (eid, e, W) where eid is an event-id, e is an event and W is a probability distribution over $\mathcal{S}$. The distribution W contains pairs of the form (σ, c), where $\sigma \in \mathcal{S}$ and $0 < c \leq 1$ is the confidence that the event e is associated with source σ; we assume $\sum_{(\sigma,c)\in W} c = 1$. A *possible world* D^* of D^p is generated by taking each event e_i in turn, and assigning it to one of the possible sources $\sigma_i \in W_i$, where $\sigma_i \in \mathcal{S}$. Thus every record $r_i = (eid_i, e_i, W_i) \in D^p$ takes the form $r'_i = (eid_i, e_i, \sigma_i)$, for some $\sigma_i \in \mathcal{S}$ in D^*. By enumerating all such possible combinations we get the complete set of possible worlds. Assuming that the distributions associated with each record r_i in D^p are stochastically independent, the probability of a possible world D^* is $\Pr[D^*] = \prod_{i=1}^{n} \Pr_{W_i}[\sigma_i]$.

Table 1. A source-level uncertain database (L) and one possible world D^* (R) showing sources and associated events (here, $\Pr[D^*] = 0.6 \times 0.3 \times 0.7 = 0.126$).

eid	event	W
e_1	a	(σ_1:0.6)(σ_2:0.4)
e_2	b	(σ_1:0.3)(σ_2:0.2)(σ_3:0.5)
e_3	c	(σ_1:0.7)(σ_3:0.3)

source	event(s)
σ_1	$(a)(b)(c)$
σ_2	$\langle\rangle$
σ_3	$\langle\rangle$

Table 2. An event-level uncertain database (L), all possible worlds for D_2^p (C) and a possible world D^* for D^p (R) containing one world each from possible worlds of every D_i^p. (here, $\Pr[D^*] = 0.126 \times 0.48 \times 0.35 = 0.021$).

	p-sequence
D_1^p	$(a:0.6)(b:0.3)(c:0.7)$
D_2^p	$(a:0.4)(b:0.2)$
D_3^p	$(b:0.5)(c:0.3)$

$\langle\rangle$	$0.6 \times 0.8 = 0.48$
(a)	$0.4 \times 0.8 = 0.32$
(b)	$0.6 \times 0.2 = 0.12$
$(a)(b)$	$0.4 \times 0.2 = 0.08$

source	possible world
σ_1	$(a)(b)(c) = 0.126$
σ_2	$\langle\rangle = 0.48$
σ_3	$\langle\rangle = 0.35$

Event-Level Uncertainty. In some cases, the 'source' of the event is known but the 'event' itself is uncertain. Consider a scenario where employees movements are tracked in a building using RFID sensors [5]. A typical relation `SIGHTING(t,`

`tID, aID)` in PEEX system [5], denotes that the RFID tag `tID` was detected by antenna `aID` at time `t`. Consequently, PEEX processes the `SIGHTING` relation to output a higher-level *uncertain* relation such as `MEETING(time, person1, person2, room, prob)`. An example tuple such as (103, 'Alice', 'Bob', 435, 0.4) in `MEETING` means that at time 103, PEEX believes that Alice and Bob are having a meeting (event) with probability 0.4 in room 435 (source) [5]; since antennae are at fixed locations, the source is certain but the event is uncertain.

A *probabilistic database* D^p is a collection of *p-sequences* $D_1^p, \ldots, D_m^p$, where D_i^p is associated with source $i \in \mathcal{S}$, $D_i^p = \langle (e_1, c_1) \ldots (e_k, c_k) \rangle$, where the events e_j are ordered by *eid* and c_j is the confidence that e_j actually occurred. The *possible worlds* semantics of D^p is as follows. For each event e_j in a p-sequence D_i^p there are two kinds of worlds; one in which e_j occurs and the other where it does not. Let $occurred = \{x_1, \ldots, x_l\}$, where $1 \leq x_1 < \ldots < x_l \leq k$, be the indices of events that occur in D_i^*. Then $D_i^* = \langle e_{x_1}, \ldots, e_{x_l} \rangle$, and $\Pr(D_i^*) = \prod_{j \in occurred} c_j * \prod_{j \notin occurred} (1 - c_j)$. The set of all possible worlds of D_i^p, denoted by $PW(D_i^p)$ is obtained by taking all possible 2^l alternatives for *occurred*, and we say $PW(D^p) = PW(D_1^p) \times \ldots \times PW(D_m^p)$. For any $D^* \in PW(D^p)$ such that $D^* = (D_1^*, \ldots, D_m^*)$, the probability of D^* is given by: $\Pr[D^*] = \prod_{i=1}^{m} \Pr(D_i^*)$.

3 Probabilistic Frequentness

For Frequent itemset mining in probabilistic databases, measures like *expected support* [1] and *probabilistic frequentness* [3] have been used. An expected support based approach for mining sequential patterns in probabilistic databases was proposed in [6]. Here, we focus on *probabilistic frequent* sequential patterns.

Definition 1. *Given a probabilistic database D^p and its set of possible worlds $PW(D^p)$, the* support probability *for a sequence s is denoted by: $\Pr_i(s) = \sum_{D^* \in PW(D^p), (Supp(s,D^*)=i)} \Pr(D^*)$, where $Supp(s, D^*)$ is the support of s in D^*. Note that $\Pr_i(s)$ is the probability that the support of s is exactly i. Further, define the* support probability distribution *(SPD) as the vector $\langle \Pr_0(s), \ldots, \Pr_m(s) \rangle$.*

Denote by $\Pr_{\geq\theta}(s) = \sum_{k=\theta}^{m} \Pr_k(s)$ the probability that the support of s is at least θ. Given D^p and two user-specified thresholds namely *support* θ, $1 \leq \theta \leq m$ and a *confidence* $\tau \in (0, 1]$, the objective is to find all *probabilistic frequent sequences (PFSes)* s s.t. $\Pr_{\geq\theta}(s) \geq \tau$ (i.e. all s with probability $\geq \tau$ of having support $\geq \theta$). Next, we show that we can obtain PFSes by *dynamic programming* (DP) for event-level uncertainty. By contrast, we show the computational intractability of finding PFSes for source-level uncertainty. We consider the fundamental question "is s a PFS", i.e. given D^p, s, θ and τ, is $\Pr_{\geq\theta}(s) \geq \tau$?

PFSes for Event-Level Uncertainty. First, we compute the probability with which a source supports a sequence s i.e. we compute $\Pr(s \preceq D_i^p)\ \forall\ i, 1 \leq i \leq m$, as done by [6]. Then, we compute $\Pr_{i,j}(s)$, for $0 \leq i, j \leq m$, which is the probability that exactly i of the first j sources support s, by DP using the recurrence:

$$\Pr_{i,j}(s) = \Pr_{i-1,j-1}(s) \cdot \Pr(s \preceq D_i^p) + \Pr_{i,j-1}(s) \cdot (1 - \Pr(s \preceq D_i^p)), \quad (1)$$

where $\Pr_{0,j}(s) = 1,\ 0 \leq j \leq m$ and $\Pr_{i,j}(s) = 0, \forall\ i > j$. Clearly, $\Pr_{i,m}(s) = \Pr_i(s)$, for all i, and we can use this to determine if s is a PFS.

PFSes for Source-Level Uncertainty. In source-level uncertainty, an event may potentially be associated to more than one sources as shown in Table 1. Note that the DP computation in Eq. 1 computes the value $\Pr_{i,j}(s)$, which does not help in this case. For example, in Table 1, event 'b' is confused between sources σ_1, σ_2 and σ_3, but it could only be associated to one of the three in a real world, which is ignored when using Eq. 1. For example, for $s = (b)$, $\Pr_{2,2}(s) = 0$, as only one of the sources can support s. However, using Eq. 1, we obtain: $\Pr_{2,2}(s) = \Pr_{1,1} \times \Pr(s \preceq D_2^p) + \Pr_{2,1} \times (1 - \Pr(s \preceq D_2^p)) = 0.3 \times 0.2 + 0 \times 0.8 = 0.06$, which is not correct. So, Eq. 1 does not work for source-level uncertainty. We further note that it is not possible to compute the value $\Pr_{i,j}(s)$. As mentioned above that $\Pr_{k,m}(s) = \Pr_k(s)$, for all k, we say that computing $\Pr_k(s)$ (i.e. the probability that exactly k sources support s) as *Exact-k-Support* problem.

Theorem 1. *Given a probabilistic database D^p, a sequence s and a number k,$0 \leq k \leq m$, computing the* Exact-k-Support *for s in D^p is $\sharp P$-complete.*

Theorem 1 is shown by reducing the problem of computing the number of perfect matchings in a bipartite graph, a $\sharp$P complete problem [8], to the Exact-k-Support problem.

4 Conclusions and Future Work

We studied uncertainty models for sequential pattern mining and discussed *probabilistic frequentness* computation for source-level and event-level uncertainties. An empirical evaluation and comparison with *expected support* in computational cost and in quality of the solution should be an interesting direction to explore.

References

1. Aggarwal, C.C. (ed.): Managing and Mining Uncertain Data. Springer (2009)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE. pp. 3–14 (1995)
3. Bernecker, T., Kriegel, H.P., Renz, M., Verhein, F., Züfle, A.: Probabilistic frequent itemset mining in uncertain databases. In: KDD. pp. 119–128 (2009)
4. Hassanzadeh, O., Miller, R.J.: Creating probabilistic databases from duplicated data. The VLDB Journal 18(5), 1141–1166 (2009)
5. Khoussainova, N., Balazinska, M., Suciu, D.: Probabilistic event extraction from rfid data. In: ICDE. pp. 1480–1482 (2008)
6. Muzammal, M., Raman, R.: Mining sequential patterns from probabilistic databases. Tech. Rep. CS-10-002, Dept. of Computer Science, Univ. of Leicester (2010), available from http://www.cs.le.ac.uk/people/mm386/pSPM.pdf
7. Suciu, D., Dalvi, N.N.: Foundations of probabilistic answers to queries. In: SIGMOD Conference. p. 963 (2005)
8. Valiant, L.G.: The complexity of computing the permanent. Theor. Comput. Sci. 8, 189–201 (1979)
9. Wikipedia: http://en.wikipedia.org/wiki/anpr — Wikipedia, the free encyclopedia (2010), {http://en.wikipedia.org/wiki/ANPR}, [accessed 30-April-2010]