DAGger: Clustering Correlated Uncertain Data

Dan Olteanu and Sebastiaan J. van Schaik

Department of Computer Science & Oxford e-Research Centre, University of Oxford

Key Features of DAGger

- It considers the **possible worlds semantics for uncertain data**
  - natural semantics for incomplete and probabilistic databases.
  - the input is a probability distribution over a set of possible worlds, whereby each world defines a set of input objects.
  - the output is equivalent to clustering within each world and defines probability distributions for objects belonging to clusters.
- It allows for **arbitrary correlations**, which are:
  - used in results to queries in probabilistic databases,
  - obtained by structuring text using Conditional Random Fields,
  - enforced by experts and learned from data in Bayesian Networks and Markov Logic Networks.

  If correlations are ignored, the output can be arbitrarily off from the true clustering result.
- It can **compute exact and approximate probabilities with error guarantees for the clustering output**.

State-of-the-art techniques (e.g. UK-means, UKmedoids, MMVar):

- do not support the possible worlds semantics,
- lack support for correlations and assume probabilistic independence,
- use deterministic cluster medoids or expected means, and
- can only compute clustering based on expected distances.

In many cases, the output is a *hard* clustering that assigns each object to one cluster, like in deterministic k-medoids or k-means.

DAGger's Approach

- The **uncertainty and correlations in the input data are represented symbolically** in a language of probabilistic events.
- **Clustering events are captured within the same formalism**.
- This formalism supports a wide range of tasks:
  - probability computation for clustering events,
  - sensitivity analysis and explanation of clustering output,
  - different clustering algorithms, e.g., k-medoids, Markov clustering.
- All clustering events are represented within one event network:
  - **Common expressions are represented only once.**
  - Yields a highly repetitive and interconnected structure due to the combinatorial nature of clustering.
  - For k-medoids and Markov clustering, the events have the same structure at each step, and at any iteration step are expressions over the events at the previous clustering iteration.
- Compute the probability of all events by **bulk-compiling an entire event network into one decision tree.**
  - Only the current root-to-leaf path of this decision tree is kept at any one time, while exploring it depth-first.
  - **Anytime approximation with error guarantees** can be achieved by exploring small fragments of the decision tree.

k-Medoids Clustering of Certain Data

1. (**Initialisation**) Initially choose an object as medoid for each cluster.
   - Given: objects $o_1, \ldots, o_n$, and clusters $C_1, \ldots, C_k$.
2. (**Assignment**) Assign object to the cluster of the *closest* medoid.
   - "closest" defined using any distance metric, e.g., Euclidean distance, Manhattan distance or Minkowski distance.
3. (**Update**) Choose new medoid for each cluster.
4. Repeat phases 2 and 3 for a number of iterations, or until fixpoint reached.

Language of Probabilistic Events

- Propositional events over independent Boolean random variables.
- Construct that can succinctly express **real values conditioned on propositional formulas**:
  - $\Phi \otimes v$ expresses that the value $v \in \mathbb{R}$ is conditioned by the formula $\Phi \in \mathbb{B}$: `if Φ then v else 0.`
  - Sums of if-then-else expressions: $\Phi_1 \otimes v_1 + \ldots + \Phi_n \otimes v_n$
  - Comparisons of such sums: $\Phi_1 \otimes v_1 + \ldots + \Phi_n \otimes v_n \leq \Psi_1 \otimes w_1 + \ldots + \Psi_m \otimes w_m$

This language allows for succinct encoding – *independently of the number of possible variable assignments* – of sums of distances from an object to any other object in a cluster, conditioned on the uncertainty of these objects.

k-Medoids Clustering of Uncertain Data

Our approach is a realisation of k-medoids clustering on uncertain data.

- It is equivalent to performing k-medoids clustering in each possible world of the input, yet avoids the explicit enumeration of possible worlds.
- The probability that an object belongs to a cluster is the sum of probabilities of those worlds in which this event occurs.
- Each object belongs to each cluster or is medoid with a certain probability.

Examples of clustering queries:

- **membership**: does a given object belong to a given cluster?
- **medoid**: is a given object the medoid of a given cluster?
- **co-occurrence**: are given objects clustered together?

Membership event $\phi^t\left[o_i \in C_j\right]$ for object o_i and cluster C_j at step $t \geq 1$:

$$\phi^t\left[o_i \in C_j\right] = \phi[o_i] \wedge \bigvee_{1 \leq a \leq n} \left(\phi^{t-1}\left[c_j = o_a\right] \wedge \left(\bigwedge_{1 \leq b \leq n, b \neq a} (\mathrm{d}(o_i, o_b) < \mathrm{d}(o_i, o_a) \rightarrow \neg\left(\bigvee_{1 \leq l \neq j \leq k} \phi^{t-1}\left[c_l = o_b\right]\right)\right)\right)$$

Medoid event $\phi^t\left[c_j = o_i\right]$ for object o_i and cluster C_j at step $t > 1$:

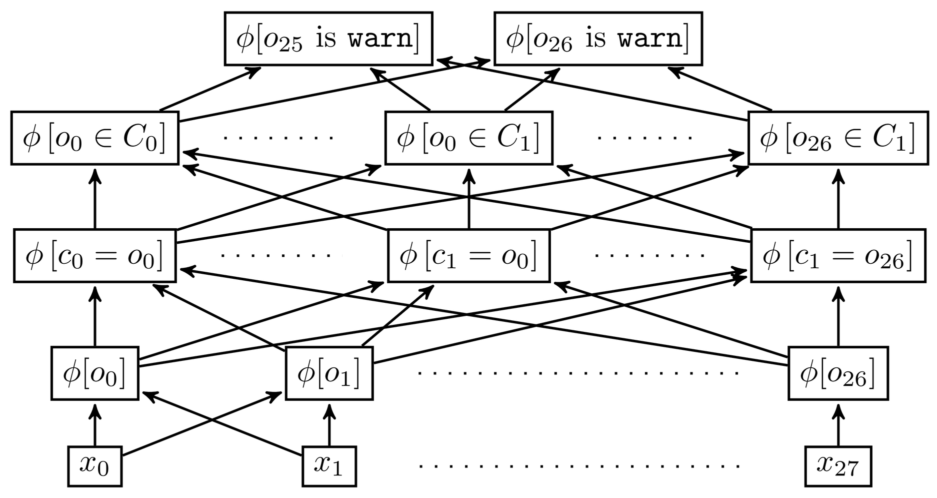
$$\Delta^t\left(o_i, C_j\right) = \sum_{a=1}^{n} \left[\phi^t\left[o_a \in C_j\right] \otimes \mathrm{d}(o_i, o_a)\right]$$

$$\phi^t\left[c_j = o_i\right] = \phi^t\left[o_i \in C_j\right] \wedge \bigwedge_{\substack{1 \leq a \leq n \\ a \neq i}} \phi^t\left[o_a \in C_j\right] \rightarrow \left(\Delta^t\left(o_i, C_j\right) < \Delta^t\left(o_a, C_j\right)\right)$$

Legend:

- $\phi[o_i]$ is the event that object o_i exists. For certain data, this event is true.
- $d(\cdot, \cdot)$ is the distance function between objects.
- $\Delta^t\left(o_i, C_j\right)$ is the total distance-sum of o_i to the objects in C_j at step t.

Exact and Approximate Probability Computation



Partial example of an **event network** with five layers encoding, highly interconnected events for clusters C_0 and C_1.

- Compilation of event network into decision tree using **Shannon expansion**: $\Phi = x \wedge \Phi|_x \vee \neg x \wedge \Phi|_{\neg x}$ This means that: $P(\Phi) = P_x \cdot P(\Phi|_x) + (1 - P_x) \cdot P(\Phi|_{\neg x})$
- If Φ is the network, then the restrictions $\Phi|_x$ and $\Phi|_{\neg x}$ are obtained by **masking** in Φ those nodes that become true or false.
- Repeated application of Shannon expansion eventually masks nodes in the network and adds the probability of the variable assignments (x or $\neg x$) to the probability mass of these nodes.
- **Approximate probability computation strategies** decide how to invest (eagerly, lazily, or hybrid) the error budget while exploring the decision tree.

Experimental Evaluation with k-Medoids Clustering of Uncertain Data

- **naive** means k-medoids in each possible world.
- types of correlations considered: **positive**, **mutex** (block-independent disjoint); **conditional independence**.