Imprecision, Diversity and Uncertainty: Disentangling Threads in Uncertainty Management

Myra Spiliopoulou, Maurice van Keulen, Hans-Joachim Lenz, Jef Wijsen,
Matthias Renz, Rudolf Kruse, and Mirco Stern

1 Introduction

The motivation of the workgroup was to couple *uncertainty management* with the notion of *change*. This encompassed *integration*, *aggregation* and *mining* upon uncertain data. Target questions were *measuring uncertainty* and *defining and measuring the quality* of uncertain data. We have left out *ambiguousness* (in the sense of unclear terms and vagueness) as an orthogonal issue. Uncertainty management in the context of reasoning (cf. among else [GS98]) has been a source of inspiration, since that area encompasses appropriate theoretical underpinnings for addressing some of the questions studied here.

2 Setting the Scene

Elaboration on uncertainty requires a context that serves as ground truth. So, under the label *"Uncertainty with respect to what?"* we have distinguished between: *Data uncertainty* caused by one or more of (1) imprecision due to metering, (2) variability due to nature and (3) variability due to sampling (cf. Examples in [Len08]), and *Model uncertainty* caused by one or more of (1) errors in specification (e.g., which distribution, overspecification of integrity constraints, etc.) and (2) parameter estimation.

The distinction between data uncertainty and model uncertainty led us to the question *"What to trust?"*, in the sense that the observer relies either on the model or on the data as ground truth. We distinguish between two fundamental cases in uncertainty management:

1. The model is true $\Longrightarrow$ If the data disagree with the model, repair the data.
2. The data are true $\Longrightarrow$ If the model disagrees with the data, revise the model.

The first case shows up, among others, when one enforces integrity constraints in a database: If some data violate the constraints, these data must be repaired. An example of the second case is concept drift in a data stream, whereupon the model must be adapted to the new data. Another example is the adaptation of a belief network with new evidence.

In practice, the observer must decide whether the model should be taken for true or rather the arriving data. Appropriate uncertainty indicators are necessary for this decision. This challenge becomes even more demanding when data

Dagstuhl Seminar Proceedings 08421
Uncertainty Management in Information Systems
http://drops.dagstuhl.de/opus/volltexte/2009/1937

uncertainty and model uncertainty co-occur. For example, the data recorded by a sensor network may contain metering errors, while the model learned from those data may itself suffer from parameter estimation errors.

3 Measuring Uncertainty

Uncertainty is not observable. To capture it, we need *uncertainty indicators*, one of which is *quality*. Scholars on data quality have identified different aspects of data quality and have even devised taxonomies of data quality indicators (cf. [Nau02] and more recently [BS06]).

Scholars on data mining and machine learning have also defined model quality indicators. For example, accuracy is a measure of classifier quality towards a ground truth. Cluster quality can be expressed as cohesion, separation and stability. The Sum of Square Errors is an example measure of cohesion, while the silhouette coefficient captures cohesion and separation [TSK04, Ch. 8].

The issue of measuring uncertainty can be seen from two perspectives: devising measures for uncertainty and designing a generic measuring process. This second perspective is important when uncertainty refers to complex *decomposable objects*, such as a model, a multidimensional data point or a dataset. One task in measuring uncertainty is thus to decompose complex objects into *atomic/non-decomposable* ones, the uncertainty of which can be measured. There is a standard repertoire for such measurements, encompassing, among others, confidence intervals for the values of variables. Then, *aggregation* is performed by propagating the atomic uncertainty measurements towards the composite object.

Following operators can be applied when measuring uncertainty: (1) *Transformation*, also known as *aggregation* in DB terminology, (2) *Conditioning – selection* or *slice*, (3) *Marginalisation – projection* or *dice* and (4) *Fusion – join.*

Applying these operators requires defining them for arbitrary types of complex objects. We take a group of clusters $\zeta = \{C_1, \ldots, C_n\}$ built by a partitioning algorithm like K-means as an easy example of a complex object: It can be intuitively decomposed into clusters/partitions. The silhouette coefficient measure computes quality for each single object x with respect to the cluster to which it was assigned and with respect to other clusters. Aggregation to the quality indicator for the whole group of clusters is computed as follows [TSK04, Ch. 8]:

$$s_{object}(x) = \begin{cases} 1 - \frac{a}{b} \ , a < b \\ \frac{b}{a} - 1 \ , a \geq b \end{cases}$$

where a is the mean of the distances between x and each other object in the cluster C containing x, and *beta* is the mean of the distances between x and each object in a cluster other than C. Then $s_{cluster}(C) = \frac{1}{|C|} \sum_{x \in C} s_{object}(x)$ so we get $s(\zeta) = \frac{1}{|\zeta|} \sum_{C \in \zeta} s_{cluster}(C)$.

For more sophisticated objects, the object decomposition and aggregation of quality indicators are less straightforward to specify. For example, cluster decomposition and aggregation of cluster quality indicators becomes more sophisticated if the clusters are allowed to overlap.

4 Exploiting Uncertainty

In some cases, uncertainty is inevitable. For example, the data delivered by a field sensor are prone to metering errors. In other cases, uncertainty can be avoided but may still be desirable. As pointed out by J.M. Keynes "It is better to be vaguely right than to be precisely wrong."

Beyond this obviously legitimate objective, it may also be preferable to be vaguely right than to be precisely right, among else if the cost of absolute certainty is too high. For example, a user issuing a query may decide to sacrifice some exactness for the advantage of getting an answer sooner. Another example is an intrusion detector: The detector can only be certain when the intrusion is complete, but its very purpose is to raise alarm in advance, so that the intrusion is prevented. This leads us to the challenge of optimizing the tradeoff between uncertainty and the cost of being certain.

5 Summarizing the challenges

We have identified the following challenges on dealing with uncertainty:

- There is uncertainty in the data and there is uncertainty in the models describing the data.
  How to decide between trusting the data and trusting the model? What uncertainty indicators can we devise to assist in this decision?
- External factors may cause changes in the population we observe, such as sensor metering errors or drifts in customer preferences.
  How to deal with data that are uncertain, when concept drift or shift is also likely, i.e. when neither the data nor the model can be trusted?
- Uncertainty refers to complex objects. There are operators that decompose them into atomic ones, upon which uncertainty can be computed. However, the specification of such operators is not straightforward and varies with the nature of the objects and with the uncertainty indicator.
  How to propagate uncertainty from atomic objects to composite ones?
- Uncertainty is sometimes inevitable and sometimes even desirable. Nonetheless, uncertainty cannot be tolerated to an arbitrary extend.
  How to trade uncertainty against the cost of being certain?

References

[BS06] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methods and Techniques.* Springer Verlag, 2006.

[GS98] Dov M. Gabbay and Philippe Smets. *Handbook of defeasible reasoning and uncertainty management systems.* Kluwer, Dordrecht, 1998.

[Len08] Hans-Joachim Lenz. Spreadsheet computation with imprecise and uncertain data. Free University Berlin, 2008.

[Nau02] Felix Naumann. *Quality-driven Query Answering for Integrated Information Systems.* Number 2261 in LNCS. Springer Verlag, 2002.

[TSK04] Pan-Ning Tan, M. Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Wiley, 2004.