A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration

Reviewed by Liyue Fan

This paper studies a “Truth Finding Problem” when integrating conflicting data regarding the same entity from different sources. The conflicting data refers to multi-valued attributes, such as cast members of a movie, or authors of a book. The challenges/obstacles are missing values or false records, beside the fact that the source quality is unknown. The paper proposed a Bayesian network to model the dependencies between underlying/latent truth, records/claims reported by various sources, and the quality of each source and to identify the latent truth as well as source quality without supervision.

Strength:
1. The authors proposed a simple Bayesian network that models the probability of an observed claim with the latent truth and the quality of reporting data source.
2. With no supervision, the latent truth can be sampled to approximately maximize the posterior probability of observed records. And the quality of data sources can be also estimated with the latent truth and observations.
3. Real data sets, i.e. books and movies, as well as synthetic dataset were used in the evaluation. All existing conflict-solving or truth finding methods were included for comparisons.

Weakness:
1. The initial distributions of the parameters, such as the quality of source and prior of latent truth, are not well explained about why beta distribution is used.
2. The Bayesian network is based on simple assumptions that may only work well when participants are trustworthy. For example, the data sources are believed not to report false data when estimating the latent truth. This can be exploited by malicious users by generating a large amount of bad data sources. Therefore, the true data records can be marked false because of low support, which in the end will undermine the quality of benevolent data sources.
3. In the evaluation, utility results such as precision and recall with synthetic data set were not reported. I suspect the performance of the Bayesian network when true data records and false data records are uniformly distributed.
Discussion:

1. This paper solves a “Truth Finding Problem”, which is closer to source rank rather than data integration. It seems to me that the proposed approach only outperforms majority voting by a bit, which can be explained as the quality of some sources (due to larger data set) can be higher than the rest sources. The proposed approach works well with real-data since missing data may be the most popular errors/conflicts in integration. But if real data and false data are randomly/uniformly generated, I don’t believe it will work well too. I didn’t benefit much from reading this paper.