HoloClean: Holistic Data Repairs with Probabilistic Inference

The paper presents new data cleaning algorithm HoloClean. Most of the data cleaning algorithms use either external data sources or some kind of learning methods. HoloClean's key point is combining these two methods to increase the accuracy of the cleaning algorithm. HoloClean can reach 90% precision and 76% recall rates and these results yields to 2x better F1 rate improvement compare to similar algorithms.

Papers introduction section explains pros and cons of using the external data while performing data cleaning. Authors also mentioned that statistical analysis methods cannot clean the data in all cases. That is why HoloClean uses external data and statistical analysis methods to improve overall data cleaning accuracy. Combination of 2 different methods increase the recall rate. Moreover, HoloClean uses initial values for indicating that there are less dirty data than the clean ones. One other good side of it is using adjustable threshold value. So, user can adjust it for either performance (runtime) or higher accuracy (precision, recall). Lastly, the authors used 4 different real data with diverse numbers of data (1K, 2M) and with different error types (duplication, identity, non-systematic and systematic errors). They also compare HoloClean with 3 similar data cleaning algorithms. In the end, HoloClean has the best recall and F1 rates out of all 4 systems and it's the best algorithm for noisy data cleaning.

There aren't many examples in the paper. Since the paper consists of complex terms and algorithms, adding more examples would simplify reading the paper. HoloClean consists of lots of steps. Each step dependent to the previous one. So, if one of the steps results aren't very good, next steps accuracy wouldn't be great. Also, authors use some steps as a black box and this black box steps affects the overall system performance. Besides, due to the number of steps, HoloClean runtime isn't very good. It might have 2x recall rate but it doesn't compensate the 3-26x slower runtime. For instance, in Food data experiment, HoloClean has lower precision and 18 times slower runtime than KATARA but it only has 2x better recall rate. Users can adjust the threshold T value for better runtime speed but, it can only halve the runtime speed while reducing the recall rate significantly.

Introduction section's first part is very good but its second part is too deep. This makes the introduction section longer and complex for readers. If we look at the algorithm itself, it outperforms other algorithms on recall rate but its precision rate is close to KATARA. I don't think 2x recall rate improvement satisfies the even 26x slower runtime. Also, with larger datasets, HoloClean takes too much time to compute. In the computation time, new values could be added into the dataset and HoloClean has to clean them later. The authors should improve runtime of the algorithm with reducing the number of steps.

2019-02-24

Baran Kaya, kayab@mcmaster.ca, 400284996