ActiveClean: Interactive Data Cleaning for Statistical Modeling

This paper presents a new algorithm for data cleaning for statistical modeling. The algorithm is called ActiveClean. It detects dirty data and uses machine learning techniques to clean the data. Since it uses learning methods, it can update its model to clean better for the next iterations. However, in the first stage, it requires people to analyze the data and to clean the data for initial learning. The aim of this algorithm is increasing correctness and efficiency of the data cleaning process.

Introduction section summarizes data cleaning methods and why it needs iterations. It also shows with graphs that some data detection methods like linear regression might not be accurate for detection. ActiveClean (AC) uses data sampling technique. However, instead of using samples from all the data, AC only gets samples from the dirty data. This method reduces the overall computation time and analyzers' burden. The authors also mention that AC is improved version of the ActiveLearning (AL). AL can only clean the tuples with null values but AC can detect and clean the incorrectly labelled data as well. AC is a dynamic system which means it updates the model to find the best loss function for optimal results. Besides, users can change the AC batch size with an input value. Users can choose larger batch size for less frequent updates and slower clean times or they can choose smaller batch size for faster cleans and frequent updates. Lastly, they tested AC with both real and syntactic scenarios/data. They also used other similar algorithms in the same experiment so that, readers can compare AC's performance.

Figure 1 shows that linear regression might not fit onto the correct results, but ActiveClean uses linear regression technique. This may cause confusion in readers. Also, for deciding the data dirtiness, people have to analyze the data. And the authors mention that they can continue analyzing or can stop whenever they want. Nonetheless, they don't mention when should analyzers stop working on the data. Is it after analysing 5 branches of data or after getting 90%+ of accuracy? Moreover, authors said that the analysts shouldn't miss certain type of data dirtiness because it could affect the model's accuracy. However, if the analysts miss some of the dirty data, how much does this affect the overall accuracy? They didn't mention about it.

There might be more examples to explain some of the complex sections but there are only few examples. For instance, authors could add an examples into the section 2. Also, if the Oracle system has better performance than the ActiveClean, is it okay to show that in the results? Generally, research papers show their algorithm as the best one, but they added Oracle in the performance comparison. I'm not sure why they did that (Because Oracle is a commercial system?). ActiveClean uses dirty data detection but this module only detects the null values and tuples that are not obey the rules. As far as I understand, this dirty data detector cannot find the tuples with misspelling or duplicated tuples. Since it cannot detect duplications and misspelling as dirtiness, it cannot show these to the analyzer to clean them. So, ActiveClean might be missing big part of the dirty data.

2019-02-24 Baran Kaya, kayab@mcmaster.ca, 400284996