Name: Bharghav Srikhakollu

CS549 – ST/Scalable Querying & ML – Spring 2022

Topic: Learning over dirty data without cleaning

https://arxiv.org/pdf/2004.02308.pdf

What is the problem discussed in the paper?

The real-world datasets are dirty and inconsistent with errors. It is due to duplicates, inconsistencies in representation of data values and violation of integrity constraints. Learning over such data will result in inaccurate models. It was understood that data scientists spend most of their time in 'cleaning and organizing data'.

There are certain approaches for data cleaning but most of the approaches are difficult and time consuming. On dirty dataset, a separate preprocessing step is present for cleaning the data and for sending the clean data to the Machine Learning algorithm. This involves significant manual effort, computational and financial resources. So, the paper talks about the possibility of eliminating the data cleaning step. The goal of the paper is to learn accurate models directly over dirty data without cleaning.

Why is it important?

Any learning over dirty data will significantly reduce the accuracy of learning. So, it is very important either to clean the data with better approaches or improve the process of learning by the ability to learn over dirty data directly. The cleaning of database is also not possible all the time since the information required to repair the errors in the database is not available and there are multiple possible clean versions that can be created for a single dirty database. Most of the times, the real-world databases prevent the relational learning algorithms from finding an accurate definition. Although there are some systems that aim to produce single probabilistic database that contain information about a subset of possible clean instances, these systems do not address the problem of duplicates.

What are the main ideas of the proposed solution for the problem?

The paper proposes a novel learning algorithm called DLearn (Dirty Learn) to learn over the inconsistent data. The approach helps to guide the learning algorithm to deal with dirty data. Data constraint types like Conditional functional dependencies (CFD) and Matching dependencies (MD) are used to guide the learning. CFD generalizes the functional dependency. MD resolves duplicates and connects entities. But each has its own drawbacks for implementation. So, the solution for the problem is producing and generalizing repairs while learning. DLearn is a bottom-up approach which is an extension to present relational learning algorithm. Most specific definition is created, and each definition is generalized to cover other examples. Generalization is done till we achieve score with more positives than negatives.

Reference citation: Learning Over Dirty Data Without Cleaning | Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data