Review of The Data Linter: Lightweight, Automated Sanity Checking for ML Data Sets [1]

Summary

This paper introduces a novel tool which is called "data linter" and is applied to machine learning (ML) datasets. This tool is a general-purpose tool that improves model quality by simplifying the time-consuming and error-prone process of data cleaning. The data linter tool analyzes ML datasets automatically, finds data issues, and offer feature transformations for a specific ML model type. The paper states how the data linter tool can help developers who are new to machine learning by teaching them and showing them how to correctly prepare data. Also, the paper provides a comprehensive taxonomy of data lints and shows that most of the datasets have at least one data lint.

Reasons to accept the paper

- The paper presents a new and novel tool which helps developer to clean and organize data for making ML models. Because of the novelty and freshness of data linter tool, it is an important contribution to the field of ML.
- The implementation of data linter tool is available as open-source. This allows others to improve the tool and encourages them for more collaboration.
- Data linter tool allows users to add new data lint detectors which makes the tool adaptable to various ML models and data.
- The educational aspect of data linter tool for new ML developers makes the paper more effective. The data linter tool explains things clearly, gives warnings, and offers suggestions to guide users for data preparation.

Reasons to reject the paper

- The paper does not provide a good comparison with existing tools or approaches in the field of data cleaning or data quality assurance. It dose not highlights the strengths and weaknesses of existing studies and their difference with data linter tool.
- It is difficult to know the importance and effectiveness of this tool because there is not any benchmarking against other methods.
- The paper evaluates the data linter only by using deep neural networks and a small set of datasets from Kaggle (600 Kaggle datasets), but it doesn't show how well the data linter works with different ML models or larger set of datasets.
- There is no discussion about potential problems in which the tool doesn't work well.
- The paper doesn't provide any feedback from users who used the tool about how easy data linter tool is to use and how practical it is in real situations.
- The paper doesn't talk about how using the data linter can affect privacy and biases in the data that is an important ethical issue.

Comments

I recommend the chairs/editors to accept this paper. This paper can have a great impact on ML field because of its contribution, open source implementation, user-extensible characteristic, and educational support, although there are some limitations that the authors should address.

Major Comments

- In the evaluation section (section 5), the authors only look at deep neural networks (DNNs). Can the data linter work well with other types of ML models? I suggest the author to evaluate the tool by using different kinds of models, not just DNNs.
- In the section 5.1, why do the authors analyze performance just by using medium-sized of datasets ("In terms of performance, for medium-sized datasets of O(100k) examples"). What about large datasets?
- In section 3.1, the authors briefly talk about scalability. I suggest the authors to have a deeper discussion about it and e.g., state what computer resources does it need? It seems a bit vague.
- In section 5.2, I suggest the authors to have a deeper analysis of false positives and false negatives which are generated by the data linter tool, because It is hard to see how reliable data linter tool is and how possible to make it better. Why you just selected 35 data sets to find out about the prevalence of issues. Provide instances about where the data linter tool generates inaccurate warning or miss potential issues.
- In section 5.1, the authors provided some examples about the educational support of data linter tool for new and matured developers in ML field, but whether this educational aspect helps users to learn the best ways for data preparation, and whether there are times when it makes a big difference in developers' understanding or not.
- In section 4, the authors provide a taxonomy of lints that is comprehensive, but I suggest the authors to make it clear by providing more examples and real-world scenarios for each lints. Also, in section 4.1, the authors mention "many of the lints described below are applicable to a range of model types", I recommend the authors to specify which lints are applicable to which ML models, because using the term "many" is a little vague now.
- The section 2 (related work) is short and does not compare data linter tool with other tools or methods well. The data linter is just compared with code linters. I recommend the authors to provide more detailed comparison that make the difference between data linter tool and other studies clear, shows the advantages of data linter tool than the others studies and shows the limitations of other studies.
- In section 6 (conclusion) the authors do not clearly summarize the main findings and contributions. I suggest the authors to mention how the data linter can be useful in practice and explain how it makes the data preparation process easier and improves model quality.
- In section 5 (evaluation), the authors talk very briefly about one limitation that exists in their study ("...but acknowledge that the validity of this analysis is limited by our own knowledge of ..."). It is better if you add a "Discussion" section to the paper and address any unexpected outcomes or limitations that you encountered during your study.

Minor comments

- I recommend the authors to use the same words consistently in the paper. E.g., sometimes the authors use "ML" and other times they write "machine learning". Keep it the same for clear understanding.
- I suggest the authors to improve the language in the paper. For example, in page 1, in introduction section, the authors can replace the phrase "when a human is the source of the data", with the phrase "when the data is generated by human".
- In section 1, page 1, the comma should be removed in the sentence: "The data linter is a tool that analyzes a user's training data and suggests ways features can be transformed to improve model quality, for a specific model type.".
- In section 4.2, page 4, the space character should be added between these two sentences: "However, some instances have a different length.Ensure that the data are materialized as expected and the model can handle data lists of varying length.".
- I recommend authors to change the location of "related work" section and add it before "conclusion and future work" section.

References

[1] Terry, M., Sculley, D., and Hynes, N. The data linter: Lightweight, automated sanity checking for ml data sets. In *Machine Learning Systems Workshop at NIPS* (2017).