Strategies for Machine Learning on heterogenous datasets

Summary:
This paper emphasis on the ways to deal with heterogenous datasets, which will often be confronted in Machine learning. The author has first classified the types of heterogeneity, and then given the solutions to each type in next section. Additionally, the author has also introduced some other related solutions such as “transfer learning”. As a conclusion, the author mentioned that researchers should analyze the problem and then carefully determine the best applicable solution.

Pros:

1. The classifications of some parts (Missing value, Distribution-based Data Irregularities, Imputation…) are really detailed.
2. The given figures can help readers to quicker and deeper understand the concepts.
3. In section “the solution of missing value”, the detailed applicable situations are given.

Cons:

1. The section “PCA and LDA” and “the solution of high dimensionality” are a bit difficult to understand.
2. “Feature-Space Remapping” is the only subsection in “Transfer Learning”, so maybe it’s better to remove the subsection title and directly describe this method.

Suggestion:

1. In section “Distribution-based Data Irregularities”, it will be better to describe given figures instead of only pasting them in the paper. Adding several sentences describing them to corresponding parts will help readers to understand.
2. Add some figures to PCA and LDA can help readers to understand.
3. Add some comparisons to the solutions under one class. For example, the author can conclude the advantage and disadvantage of Under-sampling and Over-sampling, and then gives a conclusion that in which situation, over-sampling is more suitable than under-sampling.
4. In the conclusion the author can say more about the opinion of this study and give out the possible future direction in this field.