ML1819 Research Assignment 1

Team 50

102 - “Dataset Pruning: What is the effect on Machine Learning Performance?”

Liam Greene: 15320482 - Author

Ciaran McGarry: 14321016 – Co-author

Andrew Knox: 14320552 – Contributor / Editor

**Word Count:** 987

**URL to source code repository:** https://github.com/liamogreene/CS4044-Machine-Learning-Team-50-Project-

**URL to source code repository activity:** https://github.com/liamogreene/CS4044-Machine-Learning-Team-50-Project-/graphs/contributors

j. Screenshot showing the commit activity of all students, e.g.

Dataset Pruning: What is the effect on Machine Learning Performance?

|  |  |  |  |
| --- | --- | --- | --- |
| Liam Greene Trinity College Dublin Ireland greenel3@tcd.ie | | Ciaran McGarry Trinity College Dublin Ireland cmcgarr@tcd.ie | Andrew Knox Trinity College Dublin Ireland knoxshia@tcd.ie |
|  |

# Abstract

# Introduction

Data is now seen as one of the most valuable resources in the world, even surpassing the value of oil [1]. Data has become an essential facet of the world economy, with great efforts being made to organize, categorize and work with available data. An important aspect of such work is the pruning of datasets, which involves preprocessing datasets to remove suboptimal data, such as noisy or mislabeled data, to reduce the effect ‘bad’ training examples may have on performance e.g. overfitting to outliers. This study examines the effects of dataset pruning on machine learning algorithms with the aim of identifying both advantages and disadvantages of dataset pruning to highlight situations in which dataset pruning would provide optimal results and underscore circumstances where dataset pruning would be unfavorable.

# Related Work

Dataset pruning research delves into multiple areas of study, ranging from business retail to computational linguistics. Kunz et al. note that data pruning is often a requisite of imperfect datasets. While this process improves aggregate model scores, if not carefully executed it can lead to bias and suboptimal model estimates [2]. Further research identified similar trends. The results of dataset pruning was gauged using eight benchmark tests, with five tests showing improvements, illustrating the often inconsistent nature of dataset pruning [3]. Anelia Angelova also notes ”that the learning algorithm might be better off when some training examples are discarded” [4]. Previous research points to the inconsistent nature of data pruning.

# methodology

Describe the machine learning algorithms selected and how you went about selecting appropriate values for the algorithm parameters. Present plots justifying your choices and discuss your decisions. Given the limited time you have, we do not expect a perfectly tuned system. Rather it is the critical discussion here that is important, and this should cover the major issues affecting your choices plus the level of uncertainty that your analysis indicates for the parameter choices. Also, explain and justify how you evaluate your work (e.g. chosen metrics, how training and test data was split …).

The effects of dataset pruning were examined using UCI’s Auto-Mpg Data. This dataset details the following information pertaining to automobiles:

1. mpg: continuous

2. cylinders: multi-valued discrete

3. displacement: continuous

4. horsepower: continuous

5. weight: continuous

6. acceleration: continuous

7. model year: multi-valued discrete

8. origin: multi-valued discrete

9. car name: string (unique for each instance) [5]

This dataset was chosen as it as it allowed for the study of data pruning on various variable types, including integers and strings. The dataset allowed for research into the differences that occur after the dataset pruning of variables of the same type. The dataset also contained 398 instances, providing a large sample size, while also containing missing values, which provided us with the opportunity to examine the effects of dataset pruning on imperfect datasets. Three pruning strategies were employed and compared:

1. Minimum error – Data entry errors and mistakes were removed to create an ideal dataset.
2. Grubb’s Test – statistical test to detect extreme outliers which were then removed from the dataset.
3. None – The data was not preprocessed at all

# Results & Discussion

Present, explain and discuss the results that you obtained. Include tables and figures where appropriate. Finally, answer clearly the research question. The answer must be based only on your own experiments and results. Discuss your results also under consideration of the related work. For instance, are your results confirming the results of related work or contradicting it?

# limitations & outlook

Discuss the limitations of your work, and what steps you would undertake next if you were to continue the project.

ACKNOWLEDGMENTS

This analysis was conducted as part of the 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin [2].

REFERENCES

|  |  |
| --- | --- |
| [1] | "The world’s most valuable resource is no longer oil, but data", *The Economist*,2018.[Online].Available: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. [Accessed: 29- Oct- 2018] |
| [2] | T. Kunz, S. Crone and J. Meissner, "The effect of data preprocessing on a retail price optimization system", *Decision Support Systems*, vol. 84, pp. 16-27, 2016. |
| [3] | M. Asim1, A. Rehman,, M. Idrees, “Effect of Pruning on Feature Ranking Metrics in Highly Skewed Datasets in Text Classification”, IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.10, pp. 135-149 October 2017. |
| [4] | M. A. Angelova, “Data Pruning”, California Institute of Technology, May 2004. |
| [4] | Dua, D. and Karra Taniskidou, E., “UCI Machine Learning Repository” [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2017. |