Effect of Data Pruning On Algorithm Performances\*

|  |  |  |  |
| --- | --- | --- | --- |
| Debrup Chakraborty Department of Computer Science & Statistics Trinity College Dublin Dublin, Ireland chakrabd@tcd.ie | | Viren Chhabria Department of Computer Science & Statistics Trinity College Dublin Dublin, Ireland chhabriv@tcd.ie | Aneek Barman Roy Department of Computer Science & Statistics Trinity College Dublin Dublin, Ireland barmanra@tcd.ie |
|  |

# Introduction

In the past, there have been a lot of research activities on various predictive algorithms, followed by its measure on accuracy and consequently its comparison. It is important to have an unbiased dataset as it will help the researchers to test their hypothesis in an able manner. Most of the raw datasets require cleaning i.e. removal or correction of corrupted data. After cleansing, pruning the cleaned data is required to make accurate predictions. The technique also reduces the risk of overfitting or under-fitting. As a part of the department’s machine learning group project, we tried to find out how much pruning techniques actually effect the performance of different predictive algorithms. To elaborate, if pruning can be applied on a cleansed dataset by removing the sub-optimal features and then checking the accuracy. If an algorithm (say A) on an unpruned dataset performs better than another algorithm (say B), will algorithm B perform better on the pruned data or vice-versa? This interest has been taken forward as we picked an uncleaned movie dataset from Kaggle, applied pruning techniques and ran algorithms on the pruned dataset.

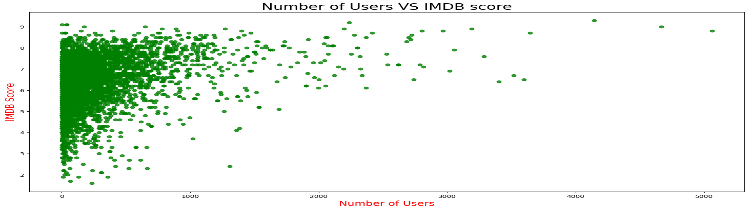


Figure 1: Plot of IMDB score v/s number of users who reviewed on a raw dataset.

# RELATED WORKs

Mladen Marovic et al [1] mentioned in their findings that the methods were not tested thoroughly because of computational complexities. Their movie dataset also lacked diversity in terms of user-ratings.

Subramaniyaswamy V. et al [2] have used multiple regression to find box-office success of a movie where they emphasized on r-value. They also achieved a better accuracy rate than previous works at SVM.

# METHODOLOGY DETAILS

## Dataset

The dataset from Kaggle is based on the imdb ratings of movies in a .csv format. It contained 5,043 entries with 28 attributes namely, imdb\_score, budget, gross, director\_name, cast\_total\_facebook\_likes, movie\_name, num\_critic-for\_reviews, num\_user\_for\_reviews etc. The visualization on the dataset provided showed that imdb score of 6.5 having the highest frequencies.

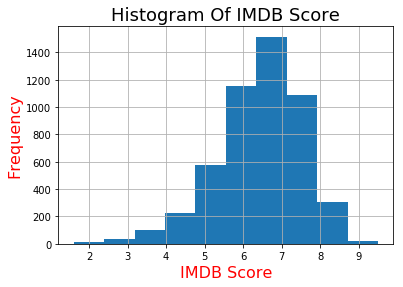


Figure 2: Frequency of IMDB Score of raw dataset

## Pre-processing

The original dataset [movie\_metadata] had 28 attributes from which 9 attributes were removed for more accurate findings, thus obtaining a cleaned dataset [processed\_df]. Some numeric data having zero or missing values were transformed using their mean values while the missing values of categorical data were classified into a missing category altogether, besides removing the duplicate values. To transform categorical data for some attributes, one-hot encoding and label encoding of scikit-learn library was used as most machine learning algorithms work on numerical data viz. country[USA] was encoded into country\_code[63], content\_rating [PG-13 ] was transformed to content\_rating\_code[ 8] etc.

## Data Pruning

The cleansed data was pruned according to the density found in the num\_user\_for\_reviews v/s imdb\_score visualization. Over a period of several incremental iterations, we selected entries where num\_user\_for\_reviews>20.

## Algorithms

We have picked three machine learning algorithms namely, Logistic Regression, Random Forest and Support Vector Machine to make a comparison of their accuracies on the pruned dataset.

*3.4.1 Logistic Regression.* The algorithm gives adjusted probability of classes and works well on large number of observations. The algorithm was chosen as it mostly have a fast prediction speed as well as training speed. We also found out that the algorithm gave best measure when c-param=1 for its range from 0.001 to 1000 with a ten time increase for each point. The following optimized problem is solved by the L1 regularized logistic regression.

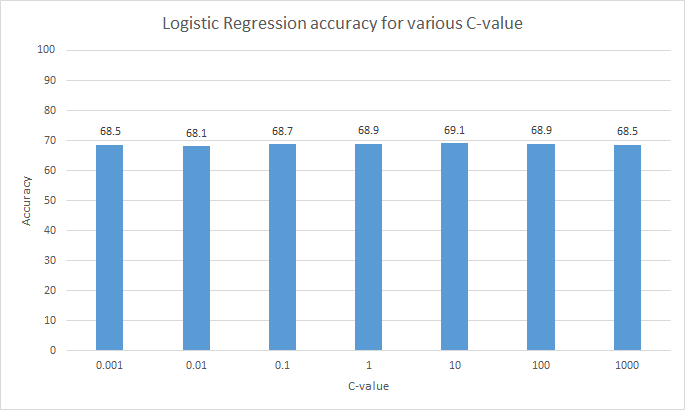


Figure 3: c-param accuracy of logistic regression

*3.4.2 Random Forest.* The choice for this algorithm is due to its effiecient handling of loads of irrelevant features of dataset unless the noise-ratio is high. The n-estimator for random forest for range [10,100,10] was found best for 40 on an unpruned data. Hence, we applied it on the pruned data as well.

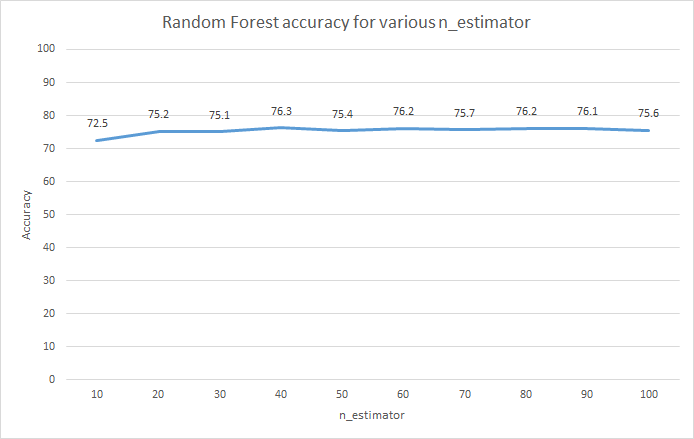


Figure 4: n-Estimator accuracy for random forest

*3.4.3 Support Vector Machine.* In our dataset, at some points there is presence of large margins which eventually help us in generalization. We ran the SVM weights to prevent overfitting on larger margins. For c-param [0.001, 0.01,0.1,10,25,50,1000], SVM was found to perform best for c-param=0.1. On the whole, this helps the test data to perform well as it doesn’t overfit the training data.

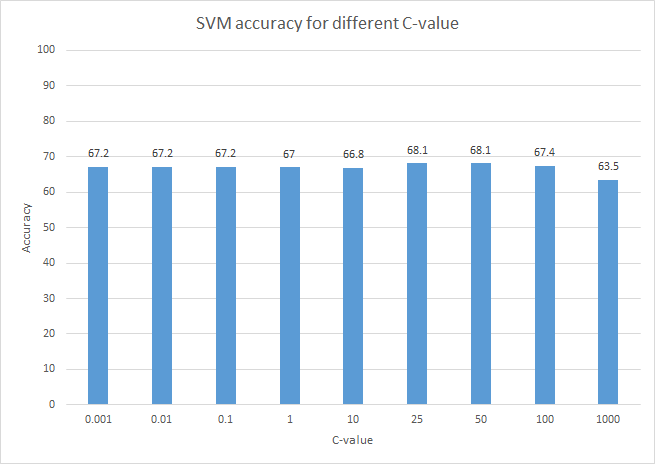


Figure 5: c-value accuracy of SVM

## Evaluation

Out of three divided train-test groups, it was found that a 20% test data would give an able accuracy measure. This was manually tested and eventually we went with the 80:20 split ratio for train-test datasets.

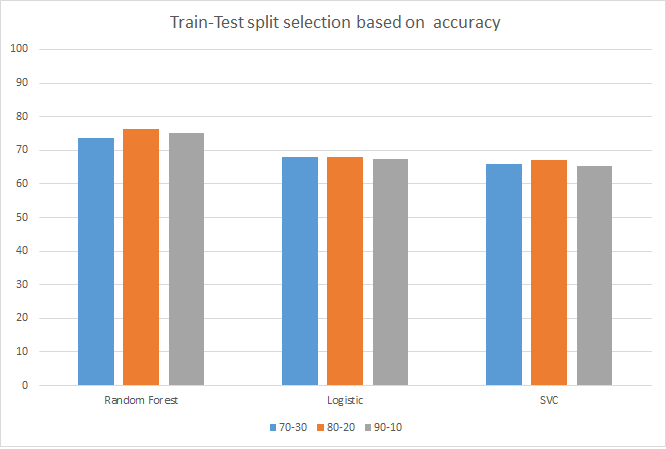
**

Figure 6: Accuracy of train-test dataset split for various algorithms

# RESULTS AND DISCUSSION

## Metrics

The models have various predictive powers which needs proper measures to evaluate the classifier. We have used accuracy score and F1-score for this.

*4.2.1 Accuracy Score.* A common metric which is the fraction of the samples correctly predicted. For a predicted value of i-th sample i.e.   and   being the respective true value, the fraction of right predictions over   may be defined as :

[4]

The mean and standard deviation of the accuracy of the three algorithms has been stated in the table below.

Table 1: **Mean and deviations of accuracies**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest | Logistic Regression | SVC |
| Mean (%) | 71.99 | 66.79 | 64.86 |
| Standard Deviation | 1.55 | 1.39 | 1.46 |

*4.2.2 F1-Score.* We selected this metric to strike a balance between precision and recall. For =1, F1 is derived from:

The mean and standard deviation of F1 scores for the three algorithms have been mentioned in the table.

Table 2: **Mean and deviations of F1 score**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest | Logistic Regression | SVC |
| Mean | 0,70 | 0.63 | 0.58 |
| Standard Deviation | 0.02 | 0.02 | 0.02 |

## Discussion

We ran 20 iterations on the dataset to check how the three algorithms responded with each iteration. For the 0th iteration, the dataset was unpruned and random forest responded the best. We used the metric measures to determine algorithm statistics and performance. For 2nd iteration, the accuracy decreased for random forest while it increased for logistic regression and decreased for SVM; however the rankings remained unchanged. On the other hand, for its 6th iteration, the accuracy for random forest increased by 3% when compared to its original accuracy, logistic regression show negligible increase whereas SVM showed 3% decrease in its accuracy. On the whole, fluctuations were seen in their respective accuracies with each iteration.

## Results

Some related literary works on movie datasets were mostly centred on random forest decision trees while some focused improving SVM accuracy. We ran unbiased analysis on the three algorithms and observed that random forest performed the best followed by logistic regression and SVM. Their rankings remain unchanged on unpruned and pruned datasets across two metric measures. However, several iterations showed some fluctuations in terms of their accuracies. To conclude, pruning of datasets didn’t affect algorithm rankings.

# Limitation and outlook

The work can be improved if analysis is run on enormous dataset or multiple datasets along with the use of multiple parameter for the three or more algorithms. The future work can also include running neural network algorithms to check their accuracies on multiple pruning techniques.

Table 1: Comparison of Accuracies of Algorithms(%)

|  |  |  |
| --- | --- | --- |
| Random Forest | Logistic Regression | SVC |
| 72.50  71.01  71.01  69.81  71.50  75.21  74.35  71.53  73.48  72.42  72.56  71.46  68.48  69.96  71.23  73.24  72.91  73.25  71.38  72.50 | 68.00  68.99  66.12  66.33  67.18  66.84  69.16  65.90  68.55  67.05  68.64  66.24  64.10  65.45  66.70  68.04  64.53  66.16  65.35  66.56 | 67.00  66.36  64.09  64.07  66.05  64.98  66.77  62.46  66.67  64.42  66.63  64.86  61.22  64.16  65.30  65.55  63.66  64.63  64.04  64.25 |

|  |  |  |
| --- | --- | --- |
| Random Forest | Logistic Regression | SVC |
| 0.71  0.70  0.69  0.68  0.70  0.73  0.74  0.70  0.72  0.71  0.71  0.70  0.67  0.68  0.70  0.71  0.71  0.72  0.69  0.71 | 0.64  0.65  0.62  0.61  0.64  0.63  0.65  0.61  0.65  0.63  0.65  0.63  0.60  0.61  0.64  0.64  0.61  0.62  0.62  0.63 | 0.60  0.60  0.56  0.56  0.60  0.58  0.60  0.55  0.60  0.58  0.58  0.58  0.53  0.56  0.59  0.59  0.57  0.58  0.58  0.57 |

Table 1: Mean and deviations of accuracies

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest | Logistic Regression | SVC |
| Mean(%) | 71.99 | 66.79 | 64.86 |
| Standard Deviation | 1.55 | 1.39 | 1.46 |

ACKNOWLEDGMENTS

This work was conducted as part of 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin [3].

REFERENCES

|  |  |
| --- | --- |
| [1] | Mladen Marovic, Marko Mihokovi ´ c, Mladen Mik ´ sa, Sini ˇ sa Pribil, and Alan Tushave. 2011. Automatic movie ratings prediction using machine learning. University of Zagreb, Faculty of Electrical Engineering and Computing Unska 3, Zagreb, Croatia |
| [2] | Subramaniyaswamy V., Viginesh Vaibhav M., Vishnu Prasad R., Logesh R. 2017. Predicting Movie Box Office Success using Multiple Regression and SVM. In Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017). IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9. |
| [3] | Joeran Beel and DouglasLeith. Machine Learning (CS7CS4/CS4404).Trinity College Dublin, School of Computer Science and Statistics. 2018. |
| [4] | Joeran Beel and Douglas Leith.Machine Learning (CS7CS4/CS4404).Trinity College Dublin, School of Computer Science and Statistics. 2018.. |
| [5] | Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171. |
| [6] | Jacques Cohen (Ed.). 1996. Special Issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996). |
| [7] | Bruce P. Douglass. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. DOI: http://dx.doi.org/10.1007/3-540-65193-429 |
| [8] | Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. DOI: http://dx.doi.org/10.1007/3-540-09237-4 |