Prediction of Abalone Ring-Age using Pruned Decision Trees and Random Forests

Data Mining and Machine Learning (ZZSC5836)

1st Waikei Lau
Faculty of Science
University of New South Wales
Sydney, Australia
z5349878@ad.unsw.edu.au

***Abstract*— The performance of Decision tree classification (CART) and ensemble learning (Random Forest) are compared in classifying abalone into four ring-age groups from eight features. 10 random-state-controlled experiments were conducted with performance on data encoding and number of estimators compared. Overfitting was managed using pre and post pruning methods; and accuracy was determined by correct classification of training and test points.**

It was found that one hot encoding of target variables significantly impacted tree classification performance. The performance differences between CART and Random Forest algorithms were small. Overall performance on the abalone dataset was approximately 62% accuracy for both methods with Random Forest demonstrating slightly higher predictive performance, and longer training times as estimators increased.

Keywords—Classification, Machine Learning, Decision Tree, Random Forest, cost complexity pruning, grid search, one hot encoding (*key words*)

I. Introduction

Predicting the age of abalone from physical measurements is often a precursor to determining laws around the recreational and industrial use of wild abalone. The determination of abalone age helps inform the management of local population sizes in the effort to prevent overharvesting and its environmental dangers. The physical determination of abalone age is similar to estimating the age of a tree in that rings are formed in the shell of the abalone as it grows at a rate of one ring per year. Determining ring-age, therefore, required drilling or cutting into the abalone shell, damaging, or killing the creature. Machine learning is an alternative to cutting into the shell, predicting abalone age by physical measurements that do not harm the animal.

This paper aims to explore Machine Learning methods that can predict the ring-age of abalone. It will focus on the implementation and performance of optimised Decision Trees and Random Forests to such a task and highlight issues related to the predictive performance of each algorithm.

II. Problem Defintiion and Algorithm

A. Task Definition

Classification Tree and Random Forest learning algorithm was applied to classify abalone into four ring-age groups from eight data features.

- Sex / nominal / -- / M, F, and I (infant)
- Length / continuous / mm / Longest shell measurement
- Diameter / continuous / mm / perpendicular to length
- Height / continuous / mm / with meat in shell
- Whole weight / continuous / grams / whole abalone
- Shucked weight / continuous / grams / weight of meat
- Viscera weight / continuous / grams / gut weight (after bleeding)
- Shell weight / continuous / grams / after being dried

These features are used to predict the variable:

- Rings / integer / -- / +1.5 gives the age in years

B. Algorithm Definition

Decision Trees (CART) are supervised learning methods used for classification and regression. The algorithm builds from the root node to minimise the Gini gain of each subsequent node by selecting the lowest Gini index node. The Gini measure is given by:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad (1)$$

The Gini index at a node is a weighted sum probability of the Gini measure for each subsequent node or leaf. The algorithm thus selects its splits to achieve the subsets that minimise Gini impurity.

The Random Forest ensemble method expands upon Decision Trees by utilising a group (‘ensemble’) of individual trees to create a **hard voting classifier**; where predictions are based on majority votes aggregated from the group. In addition, the random sampling of the training set is done with replacement (bootstrap aggregating or ‘bagging’) and the node splitting criteria is based on a random subset of features.

Of the two learning algorithms, Decision Trees are especially prone to overfitting and methods are typically employed to manage complexity and the growth in nodes. **Pre-pruning** or early stopping, as the name implies, manages overfitting by stopping the tree-building process early. It is conducted prior to the construction of a Decision Tree or Random Forest model, by a cross-validated grid search. It iterates through the hyperparameter grid and fits models with the parameters in the grid. Each parameter is assigned a score based on model performance and the parameter with optimal score is selected. **Post-pruning** and Cost Complexity Pruning provides an alternative to manage overfitting. A cost complexity parameter ccp_alpha is used to regularise tree growth by applying a penalty to the number of terminal nodes. Effective alphas are selected to minimise the final tree score according to the equation:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right| \qquad (2)$$

III. Experimental Evaluation

Model efficiency and performance was evaluated based on accuracy of predicted outcomes versus actual outcomes.

Estimators in Scikit-learn (such as Decision Tree Classifier and Random Forest Classifier) have a score method equal to the mean accuracy on the given test data and is equivalent to the accuracy classification score. Both optimised and un-optimised models were assessed on model score with the best performing option selected.

A. *Methodology*

All Decision Tree and Random Forests estimators were constructed and equally assessed on 10 random-state subsamples and assessed through stages of data extraction, pre-processing, splitting, fitting, pruning, prediction, analysis, and visualisation. All estimators were initialised against a 1 percent **minimum sample leaf ratio** to prevent overfitting.

1. Pre-processing – Target values for ring-age were classified into four classes and was one-hot-encoded with abalone gender (Male / Female / Infant). Binary data columns created:

    a. Male
    b. Infant
    c. Female
    d. Class 1: 0-7 years
    e. Class 2: 8-10 years
    f. Class 3: 11-15 years
    g. Class 4: Greater than 15 years

2. Splitting – randomised 60/40 train test splits.

3. Fitting – An unrestricted full-depth Decision Tree and Random Forest was fitted (as reference) with test and training accuracy recorded. The Random Forest was fitted on an increasing number of estimators as per the formula: $estimators = 2^{experiment\ no.} + 1$. This ensured that a wide spectrum of estimators was assessed for predictive performance.

4. Pruning – reference estimators informed the pre-pruning parameter search using the model depth as maximum range for max depth parameter.

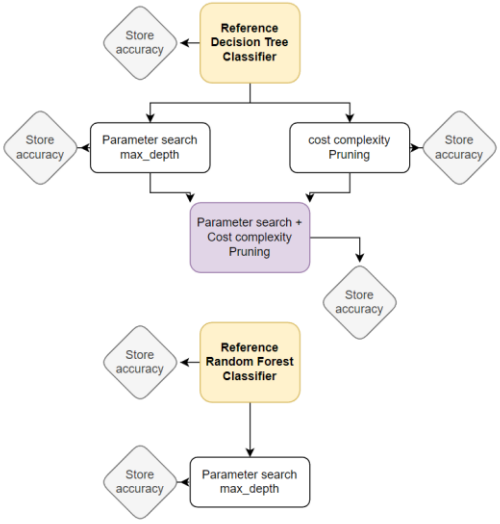

Store accuracy
Reference Decision Tree Classifier
Store accuracy
Parameter search max_depth
cost complexity Pruning
Store accuracy
Parameter search + Cost complexity Pruning
Store accuracy
Store accuracy
Reference Random Forest Classifier
Store accuracy
Parameter search max_depth

Figure 1 Decision-Tree and Random Forest estimators

    a. Five models were run in Decision Tree classification to investigate the best estimator against overfitting:

        i. Traditional Decision Tree
        ii. GridSearchCV Pruning
        iii. Cost Complexity Pruning
        iv. Applying Cost complexity pruning to the estimator in ii (GridSearchCV).
        v. Applying GridSearchCV for max depth on the estimator from iii (Cost complexity pruning).

    b. Two models were run in Random Forest classification; however, no post pruning was applied as the samples used are bootstrapped and Random Trees use random features and majority voting to mitigate overfitting. The estimators were:

        i. Traditional Random Forest
        ii. GridSearchCV Pruning

5. Prediction – all estimators were ranked according to model score measuring mean accuracy on the given test data.

B. *Results*

- Initial analysis of the data illustrates a potential class imbalance issue with Class 4 being underrepresented by 6% of samples, while Class 2 is overrepresented accounting for 45% of samples. Essentially underrepresented data induces sparsity where there may not be enough data to produce a meaningful algorithmic split for Class 4 [1].

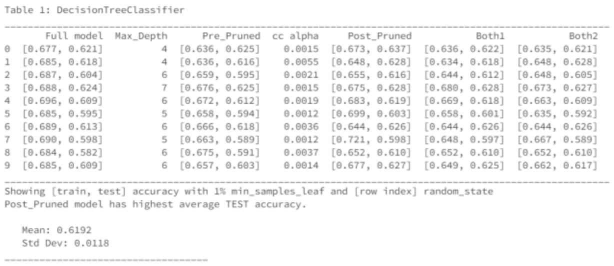
Class 1	Class 2	Class 3	Class 4
839	1891	1185	261
20%	45%	28%	6%

Table 1 Class imbalance and weightings

- One-Hot-Encoded (multilabel) target (dependent) values (d, e, f, g) impaired model performance and accuracy on the abalone dataset. The optimal mean test accuracy of Decision Tree estimators was 0.6113 with one hot encoding versus 0.6192 without. Random Tree model performance were significantly impaired by encoding with mean test performance of 0.5319 with encoding and 0.6268 without. As such it was determined that the optimal methodology would be to remove multilabel model and proceed with multiclass labels upon which the rest of this report were sampled.

  One-hot-encoding features also had statistically insignificant impacts on accuracy depending on the model. It was found to both decrease the accuracy of Random Forest Classification and improve the accuracy of Decision Tree Classifiers.

- Among Decision Tree models, the optimal Decision Tree Classifier after 10 experiments utilised cost complexity pruning. The mean test accuracy was 0.6192 with a standard deviation of 0.0118.



```
Table 1: DecisionTreeClassifier

------------------------------------------------------------------------------------------------------
       Full model  Max_Depth      Pre_Pruned  cc alpha     Post_Pruned           Both1           Both2
0  [0.677, 0.621]          4  [0.636, 0.625]    0.0015  [0.673, 0.637]  [0.636, 0.622]  [0.635, 0.621]
1  [0.685, 0.618]          4  [0.636, 0.616]    0.0055  [0.648, 0.628]  [0.634, 0.618]  [0.648, 0.628]
2  [0.687, 0.604]          6  [0.659, 0.595]    0.0021  [0.655, 0.616]  [0.644, 0.612]  [0.648, 0.605]
3  [0.688, 0.624]          7  [0.676, 0.625]    0.0015  [0.675, 0.628]  [0.680, 0.628]  [0.673, 0.627]
4  [0.696, 0.609]          6  [0.672, 0.612]    0.0019  [0.683, 0.619]  [0.669, 0.618]  [0.663, 0.609]
5  [0.685, 0.595]          5  [0.658, 0.594]    0.0012  [0.699, 0.603]  [0.658, 0.601]  [0.635, 0.592]
6  [0.689, 0.613]          6  [0.666, 0.618]    0.0036  [0.644, 0.626]  [0.644, 0.626]  [0.644, 0.626]
7  [0.690, 0.598]          5  [0.663, 0.589]    0.0012  [0.721, 0.598]  [0.648, 0.597]  [0.667, 0.589]
8  [0.684, 0.582]          6  [0.675, 0.591]    0.0037  [0.652, 0.610]  [0.652, 0.610]  [0.652, 0.610]
9  [0.685, 0.609]          6  [0.657, 0.603]    0.0014  [0.677, 0.627]  [0.649, 0.625]  [0.662, 0.617]
------------------------------------------------------------------------------------------------------
Showing [train, test] accuracy with 1% min_samples_leaf and [row index] random_state
Post_Pruned model has highest average TEST accuracy.

   Mean: 0.6192
   Std Dev: 0.0118
-----------------------------------
```

Figure 2 Accuracy scores of Decision Tree Models

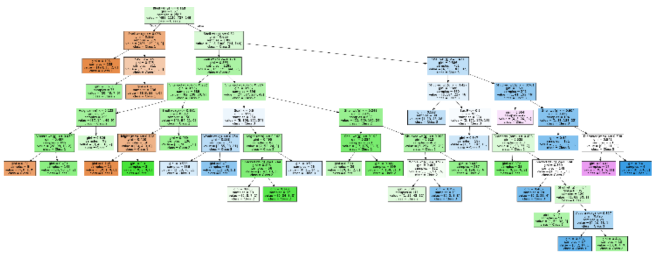


Figure 3 Decision Tree Diagram (C.C. Pruned)

a. Figure 3 can be translated into If-Then rules for a few selected nodes:

- *if* Shell weight is ≤ 0.113g *AND* ≤ 0.056g, *then* → predict Class 1 (0-7) years.
- *if* Shell weight is less than 0.113g *AND* > 0.056g *AND* is *not* infant, *then* → predict Class 2 (8-10) years
- *if* Shell weight is less than 0.113g *AND* > 0.056g *AND is* infant, *then* → predict Class 1 (8-10) years

b. GridSearchCV optimised max depth to an average of 5.5 for model ii. As evidenced from figure 4 below, higher max depth increases the separation between training and test accuracies and caused the model to overfit to the training data.

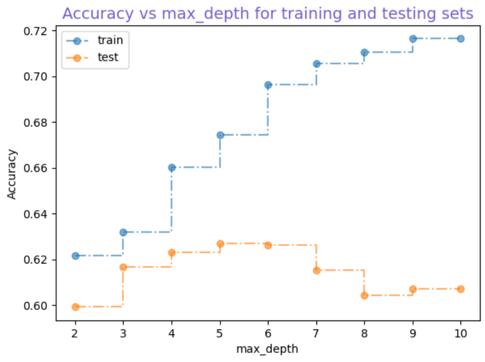


Figure 4 Max Depth vs Accuracy

c. Optimal Cost Complexity alphas averaged around 0.00236. As alpha increased to 0.002, test prediction accuracy improved while training prediction accuracy declined, reducing the overfit to training data. This continues until 0.005 when accuracy begins to decline for both testing and training samples.

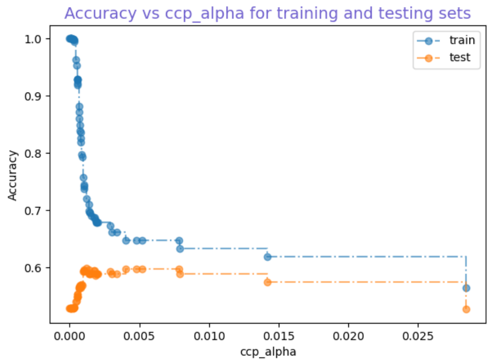


Figure 5 Ccp Alpha vs Accuracy

d. The confusion Matrix indicates there are some class imbalance issues resulting in most Class 4 samples being misclassified to Class 3. Another notable observation is 188 class 3 samples being misclassified to Class 2 and 157 Class 2 being misclassified to Class 3. This suggests there is high correlation in the features for Class 2, Class 3, and Class 4, that is exaggerating the class imbalance issue. Lastly, the F1 score of 0.62 suggests that over 1 in 3 predictions are incorrect.

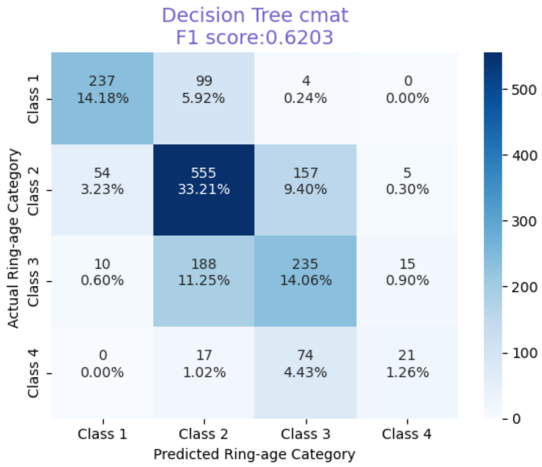
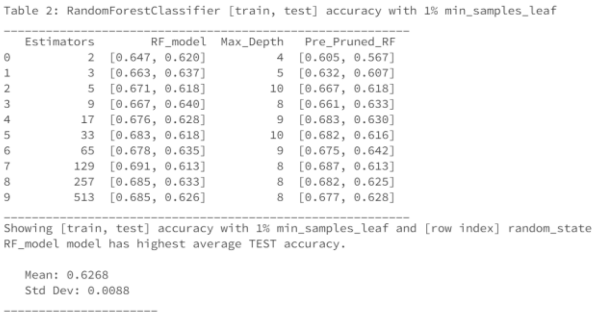


Figure 6 Decision Tree Confusion Matrix (C.C. Pruned)

- Of the two Random Forest models, the unpruned Random Forest Classifier was found to perform best with a mean test accuracy of 0.6268 and a standard deviation of 0.0088.



```
Table 2: RandomForestClassifier [train, test] accuracy with 1% min_samples_leaf

------------------------------------------------------------
   Estimators         RF_model  Max_Depth    Pre_Pruned_RF
0           2  [0.647, 0.620]          4  [0.605, 0.567]
1           3  [0.663, 0.637]          5  [0.632, 0.607]
2           5  [0.671, 0.618]         10  [0.667, 0.618]
3           9  [0.667, 0.640]          8  [0.661, 0.633]
4          17  [0.676, 0.628]          9  [0.683, 0.630]
5          33  [0.683, 0.618]         10  [0.682, 0.616]
6          65  [0.678, 0.635]          9  [0.675, 0.642]
7         129  [0.691, 0.613]          8  [0.687, 0.613]
8         257  [0.685, 0.633]          8  [0.682, 0.625]
9         513  [0.685, 0.626]          8  [0.677, 0.628]
------------------------------------------------------------
Showing [train, test] accuracy with 1% min_samples_leaf and [row index] random_state
RF_model model has highest average TEST accuracy.

   Mean: 0.6268
   Std Dev: 0.0088

----------------------
```

Figure 7 Accuracy of Random Forest models

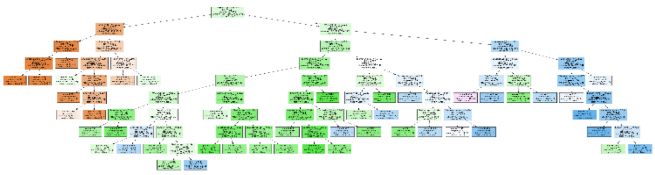


Figure 8 Single estimator from a 513 tree Forest

a. Figure 8 can be translated into If-Then rules for a few selected nodes:

- *if* Shell weight is > 0. 117g *AND* whole weight is > 1.226g *AND* ≤ 1.39g *AND* shucked weight > 0.565g *AND* shell weight ≤ 0.359g *then* → predict Class 2 (8-10) years.
- *if* Shell weight is > 0. 117g *AND* whole weight is > 1.226g *AND* ≤ 1.39g *AND* shucked weight > 0.565g *AND* shell weight > 0.359g *then* → predict Class 3 (11-15) years.
- *if* Shell weight is > 0. 117g *AND* whole weight is > 1.226g *AND* ≤ 1.39g *AND* shucked weight ≤ 0.565g *AND* is *not* female *then* → predict Class 4 (15+) years.

b. From figure 7, as number of estimators increased there was no significant improvement in predictive performance beyond attributable random variation. The computational speed was significantly longer from experiment 7 onwards due to the higher number of estimators.

c. Interestingly, GridSearchCV optimised the max depth parameter in Random Forest models to an average of 7.9. This is higher than the average Decision Tree optimised value of 5.5. Specifically, the increase in depth may be attributed to the increased randomness requiring greater average depth to fit to the training data.

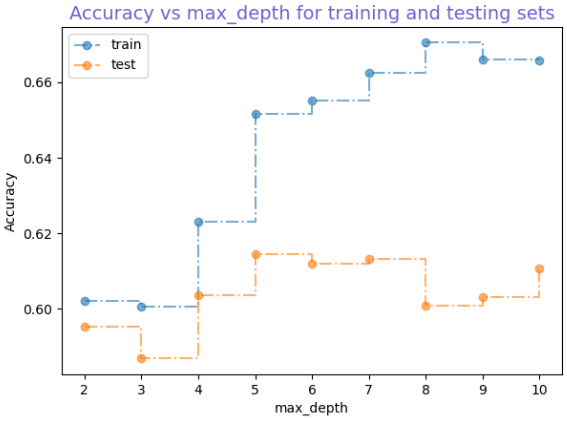


Figure 9 Max Depth Accuracy of Random Forest

d. The Random Forest Confusion matrix shows that almost all Class 4 samples were incorrectly predicted to Class 3. This illustrates that Random Forests are more susceptible to class imbalance issues which can be addressed with under/oversampling and class weighting. Again, the model was able to predict most Class 1 samples correctly but struggled to differentiate Class 3 from Class 2. The F1 Score of 0.6 indicates that the model misclassified around 2 out of 5 predictions.

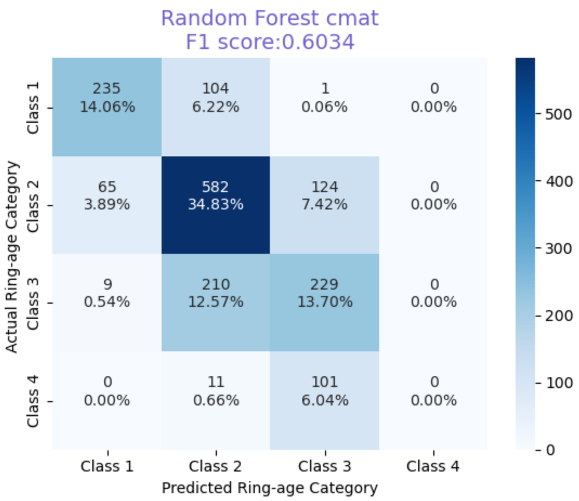


Figure 10 Confusion Matrix for Random Forest

C. Discussion

On the abalone dataset, the predictive performance of optimised Decision Tree and Random Forest Classification is similar. Both methods are not robust to class imbalance issues and do perform better when the dataset is multiclass rather than multilabel (one-hot-encoded).

The issue of one-hot-encoding is due to the tendency of Tree-based models to “pick the feature to split on based on how well that splitting the data on that feature will "purify" it. If we have a lot of levels, only a small fraction of the data (typically) will belong to any given level, so the one-hot encoded columns will be mostly zeros. Since splitting on this column will only produce a small gain, tree-based algorithms typically ignore the information in favour of other columns” [2]. This was the demonstrated problem in abalone dataset where Class 4 represented 6% of the subsample.

A substantial improvement in Random Forest predictive power was gained by moving from a multilabel (one-hot-encoded) learning task to a multiclass learning task where y labels were combined into one numerically distinct variable separating the age classes.

Random Forests in particular, were demonstrated to be particularly susceptible to class imbalance and future work should ensure class balancing techniques are carried out for such methods. Synthetic Minority Oversampling Technique (SMOTE) or Over-Under-Bagging (or ‘Balanced Random Forests) could be used to oversample Class 4 – the under-represented class – and under sample Class 3 as part of the fitting phase [3][4]. More broadly, class weighting methods may be used during the sampling phase to obtain a more balanced sample [5].

On accuracy alone, the Random Forest did outperform the Decision Tree with a slightly higher test accuracy of 0.6268, however the margin was not a statistically significant.

In conclusion, the use of machine learning on the abalone dataset has shown that improvements can still be made to model performance. However, Decision Trees and Random Forests are a promising way to decern abalone age from physical features and their use in supporting environmental initiatives should be considered in future projects.

IV. Future Work

Further research can be focused on ensemble techniques in combination with a range of Machine Learning algorithms. For instance, convolutional neural networks could help train future models to improve the classification performance by combining image processing with physical measurements of creature size and weight. This introduces new dimensionality to the data and can go a long way to mitigating the class imbalance issue.

Lastly, dimensionality reduction techniques such as Principal component analysis would be well suited to the abalone dataset which exhibits high inter feature correlations particularly for size and weight measurements. PCA supplemented with clustering techniques such as DBSCAN in ensemble with deep neural nets and Random Forests may decompose and the dataset more effectively and produce a better hard voting classifier. [6]

References

[1] R. Ravi, “One-hot encoding is making your tree-based ensembles worse, here’s why?,” towardsdatascience.com, https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769 (accessed June 10, 2022).

[2] D. Martin, “Are you getting burned by one-hot encoding?” kiwidamien.github.io, https://kiwidamien.github.io/are-you-getting-burned-by-one-hot-encoding.html (accessed June 10, 2022).

[3] J. Brownlee, “SMOTE for imbalanced classification with Python,” machinelearningmastery.com https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (accessed June 16, 2022).

[4] J. Brownlee, “Bagging and Random Forest for imbalanced classification,” machinelearningmastery.com https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/ (accessed June 16, 2022).

[5] B. Hussain, K. Huh, H. Wing, E. Chan, and S. Patanwala, “Surviving in a Random Forest with imbalanced datasets,” medium.com, https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb (accessed June 16, 2022).

[6] C. Zvi, and A. Schclar, “Ensemble classification via Kernel-PCA dimensionality reduction,” The Academic College of Tel Aviv-Yaffo School of Computer Science http://www.cs.mta.ac.il/staff/Alon_Schclar/pdf/Chen%20Zvi.pdf (accessed June 14, 2022).