Predicting the Sex of Abalone

Samuel Edeh
Department of Computer Science
Lewis University
Romeoville, IL, USA
samueledeh@lewisu.edu

***Abstract*—Much work has been done on predicting the age of abalone from physical measurements. However, these efforts have faced challenges because the abalone dataset is highly overlapped. A useful alternative to predicting age is to predict the sex of abalone. Being able to predict the sex of abalone with substantial confidence and avoid some of the problems due to the overlapped data is certainly good news to conservationists who are often faced with the task of protecting certain endangered sex of abalone.**

Keywords—Abalone; classifiers; precision; conservation; ensemble

I. Objective

The purpose of this paper is to predict the sex of abalone (whether infant, male or female) given various descriptive attributes. Abalones are edible, commercially valuable sea snails that grow on reefs. Measuring each abalone before removing it from the reef and determining the abalone's sex from the measurement will provide a useful conservation tool to allow divers to selectively harvest abalone based on sex.

II. Description of The Dataset

The Abalone Dataset, available from the UCI machine learning repository [1], contains the physical measurements of abalones. The dataset was collected with the goal of attempting to predict the age of abalone from physical measurements. In this paper, I take a different approach and instead attempt to predict the sex of abalone from physical measurements.

Each abalone or row in the dataset contains a categorical attribute (sex), 7 continuous attributes (length, diameter, height, whole weight, shucked weight, viscera weight and shell weight) and an integer attribute (number of rings). There are 4177 rows in the dataset with no missing values. The dataset is claimed to be "highly overlapped" [2, p. 5] likely due to the high similarity between the male and female classes as shown in the summary statistics in Table 1 below.

TABLE I. Dataset Summary

Attributes	Infant (N=1342)	Male (N=1527)	Female (N=1307)
Length	0.43±0.11	0.56±0.10	0.58±0.09
Diameter	0.33±0.09	0.44±0.08	0.45±0.07
Height	0.11±0.03	0.15±0.03	0.16±0.04
Whole weight	0.43±0.29	0.99±0.47	1.05±0.43
Shucked weight	0.19±0.13	0.43±0.22	0.45±0.20
Viscera weight	0.09±0.06	0.22±0.11	0.23±0.10
Shell weight	0.128±0.09	0.28±0.13	0.30±0.13
Rings	8±3	11±3	11±3

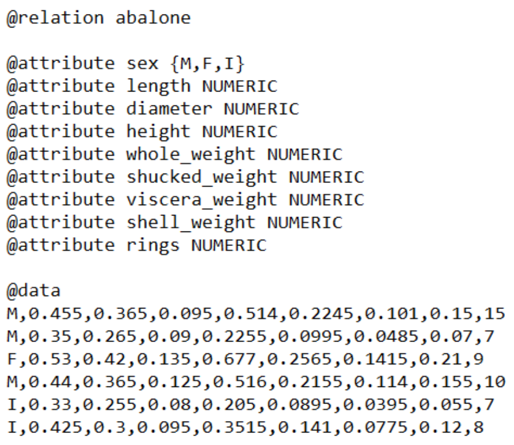
Values are presented as mean±standard deviation

A. Data Processing

I used the Weka toolkit in this prediction exercise. Weka's preferred data format is the Attribute-Relation File Format (ARFF). An ARFF file is an ASCII text file containing the name of the relation, a list of the attributes and their data type and the data itself provided as rows in a comma-delimited format [3, Ch. 9, p. 161].

A sample of the abalone ARFF file used in Weka appears in Fig. 1. The first 3133 rows of the dataset were used for training and the last 1044 were used for testing.

Fig. 1. A sample of the abalone dataset in ARFF format

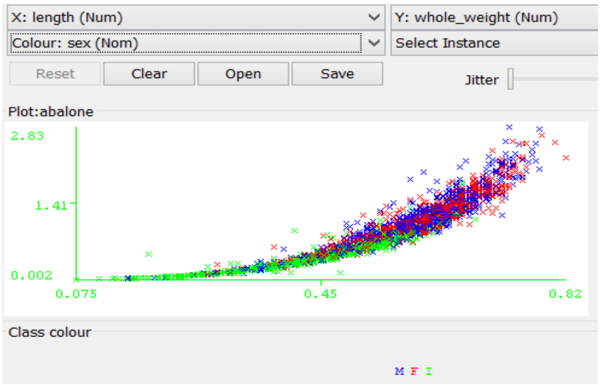


```
@relation abalone

@attribute sex {M,F,I}
@attribute length NUMERIC
@attribute diameter NUMERIC
@attribute height NUMERIC
@attribute whole_weight NUMERIC
@attribute shucked_weight NUMERIC
@attribute viscera_weight NUMERIC
@attribute shell_weight NUMERIC
@attribute rings NUMERIC

@data
M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15
M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7
F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9
M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10
I,0.33,0.255,0.08,0.205,0.0895,0.0395,0.055,7
I,0.425,0.3,0.095,0.3515,0.141,0.0775,0.12,8
```

B. Visualizing the Data

Fig. 2 below is a power-law graph relating length to whole weight with the male (blue) and female (red) classes overlapping each other. Power-law graphs of other numeric attribute pairs follow a similar pattern. This visualization provides further support that male and female classes are highly overlapped and will be difficult to separate.

Fig. 2. Visualization of the overlap of attributes of the male/female classes



III. Data Mining Process

The dataset contains both numeric and categorical attributes and, therefore, only algorithms that can handle both types of attributes can be used. Regression is automatically disqualified since it works exclusively with numeric values. Clustering would just discovers clusters and “the overall distribution pattern of the dataset” [4, p. 22] but would not produce actual predictions. “Association Rule Mining algorithms operate on a data matrix (e.g. customers × products) to derive rules” [5, p. 1]. They are used in scenarios, not available here, where “it is desirable to discover the important associations among items such that the presence of some items in a transaction will imply the presence of other items in the same transaction” [4, p. 7].

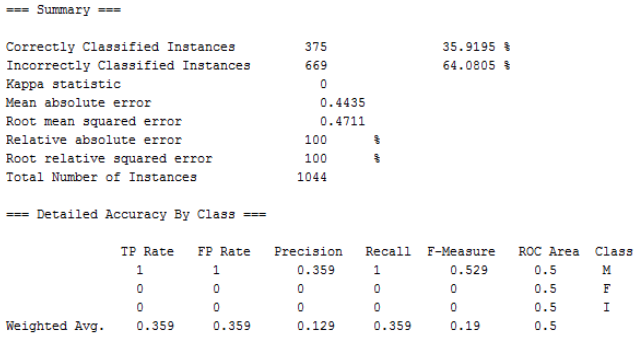
In contrast, classifiers find “the common properties” among a set of instances in a dataset and “classifies them into different classes” [4, p. 19]. The task of classifying an abalone as infant, male or female is a natural fit for classifiers. I would therefore use classification algorithms for this prediction exercise.

A. Establishing Baseline Performance

ZeroR is a simple classifier often used as a lower bound on performance. It predicts the most common class (or median in the case of numeric values) essentially testing “how well the class can be predicted without considering other attributes” [6, p. 83]. Here, male is the most common class occurring 36.6% of the time in the training set compared to infant and female which occurred 32.1% and 31.3% of the time, respectively.

Because ZeroR predicts the most common class, in this case male, the model is expected to yield 0% recall and precision for infant and female and 100% precision and recall for male. The results of ZeroR are shown in Fig. 3 and are consistent with this observation. The results also show 35.9% overall accuracy for ZeroR.

Fig. 3. Lower bound predictions based on ZeroR



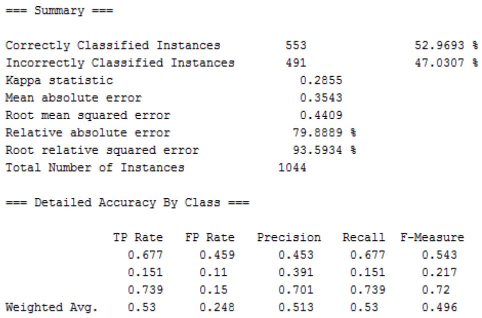
```
=== Summary ===

Correctly Classified Instances         375               35.9195 %
Incorrectly Classified Instances       669               64.0805 %
Kappa statistic                          0
Mean absolute error                      0.4435
Root mean squared error                  0.4711
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1044

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.359     1         0.529      0.5      M
                 0         0          0         0         0          0.5      F
                 0         0          0         0         0          0.5      I
Weighted Avg.    0.359     0.359      0.129     0.359     0.19       0.5
```

The most simple and worst performing classifier is of course random guess. With 3 possible classes for sex, the probability of guessing any particular sex correctly in a random guess is approximately 33.3%. My goal is that the learning models generated in this prediction exercise perform substantially better than random guess and ZeroR.

B. Classifier Parameters

I explored with 2 algorithms based on decision trees: J48 and Random Forest. I also explored with IBk which is based on k-Nearest Neighbor (k-NN). In the case of J48, I used the following parameters: *binarySplits* = FALSE, *confidenceFactor* = 0.25, *reducedErrorPrunning* = TRUE, *minNumObj* = 2 and *unpruned* = FALSE. The results of J48 are shown in Fig. 4.

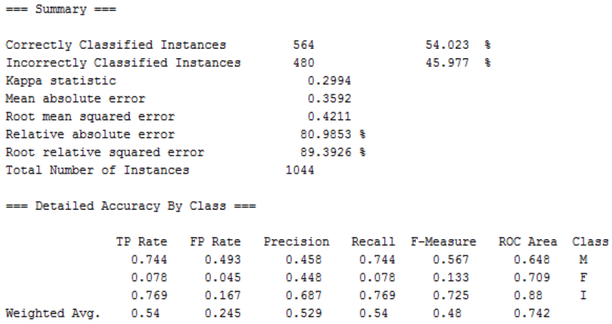
Fig. 4. J48 performance on predicting sex of abalone



```
=== Summary ===

Correctly Classified Instances         553               52.9693 %
Incorrectly Classified Instances       491               47.0307 %
Kappa statistic                          0.2855
Mean absolute error                      0.3543
Root mean squared error                  0.4409
Relative absolute error                 79.8889 %
Root relative squared error             93.5934 %
Total Number of Instances             1044

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.677     0.459      0.453     0.677     0.543      0.619    M
                 0.151     0.11       0.391     0.151     0.217      0.667    F
                 0.739     0.15       0.701     0.739     0.72       0.839    I
Weighted Avg.    0.53      0.248      0.513     0.53      0.496      0.706
```

In the case of the Random Forest algorithm, I used the following parameters: *debug* = False, *maxDepth* = 2, *seed* = 1, *numFeatures* = 4 and *numTrees* = 20. The results of the Random Forest algorithm are shown in Fig. 5.

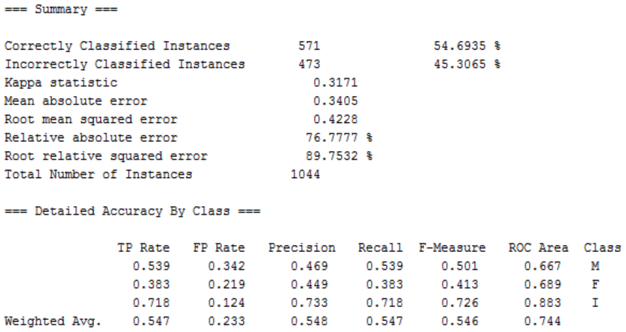
Fig. 5. Random Forest performance on predicting sex of abalone



```
=== Summary ===

Correctly Classified Instances         564               54.023  %
Incorrectly Classified Instances       480               45.977  %
Kappa statistic                          0.2994
Mean absolute error                      0.3592
Root mean squared error                  0.4211
Relative absolute error                 80.9853 %
Root relative squared error             89.3926 %
Total Number of Instances             1044

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.744     0.493      0.458     0.744     0.567      0.648    M
                 0.078     0.045      0.448     0.078     0.133      0.709    F
                 0.769     0.167      0.687     0.769     0.725      0.88     I
Weighted Avg.    0.54      0.245      0.529     0.54      0.48       0.742
```

In the case of IBk, I used the following parameters: *KNN* = 15, *crossValidate* = False, *debug* = False, *distanceWeighting* = No Distance weighting, *meanSquared* = False, *nearestNeighborSearchAlgorithm* = LinearNNSearch, and *windowSize* = 0. The results of IBk are shown in Fig. 6.

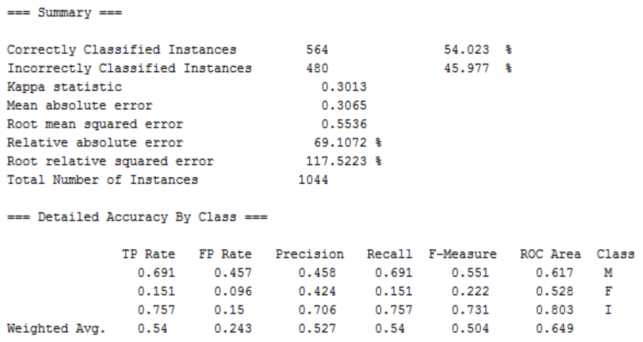
Fig. 6. IBk performance on predicting sex of abalone



```
=== Summary ===

Correctly Classified Instances         571               54.6935 %
Incorrectly Classified Instances       473               45.3065 %
Kappa statistic                          0.3171
Mean absolute error                      0.3405
Root mean squared error                  0.4228
Relative absolute error                 76.7777 %
Root relative squared error             89.7532 %
Total Number of Instances             1044

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.539     0.342      0.469     0.539     0.501      0.667    M
                 0.383     0.219      0.449     0.383     0.413      0.689    F
                 0.718     0.124      0.733     0.718     0.726      0.883    I
Weighted Avg.    0.547     0.233      0.548     0.547     0.546      0.744
```

Finally, I combined all three algorithms using the Vote ensemble method. I set the parameters of each algorithm as already described. For the Vote algorithm, I used the following parameters: *combinationRule* = Majority Voting, *debug* = False and *seed* = 1. The results of the ensemble method are shown in Fig. 7.

Fig. 7. Vote ensemble performance on predicting sex of abalone



```
=== Summary ===

Correctly Classified Instances         564               54.023  %
Incorrectly Classified Instances       480               45.977  %
Kappa statistic                          0.3013
Mean absolute error                      0.3065
Root mean squared error                  0.5536
Relative absolute error                 69.1072 %
Root relative squared error            117.5223 %
Total Number of Instances             1044

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.691     0.457      0.458     0.691     0.551      0.617    M
                 0.151     0.096      0.424     0.151     0.222      0.528    F
                 0.757     0.15       0.706     0.757     0.731      0.803    I
Weighted Avg.    0.54      0.243      0.527     0.54      0.504      0.649
```

IV. Results

From Fig. 4-7, IBk achieved the best overall accuracy of 54.7% followed by Random Forest and J48 with overall accuracy of 54.0% and 53.0%, respectively. The overall accuracy of the algorithms seems mediocre and is likely due to the highly overlapped nature of the male and female classes. The ensemble method did not improve performance.

Beyond accuracy, high precision is essential. For a diver to rely on the prediction from the models to harvest abalones on the basis of sex the diver needs to be confident that the models substantially identify the particular sex being predicted. For e.g., the precision for both male and female classes is <50% suggesting tepid confidence for prediction of those classes. However, the precision for infant is relatively high at ≥68.7% and provides greater confidence for prediction of the infant class. The recall and F-measure for the infant class are also high indicating the algorithms are able to distinguish infant abalones from male and female abalones.

V. Conclusion

A. What I learned from the results

The highly overlapped nature of the male and female classes curtailed the accuracy yields for the male and female classes. Nevertheless, the algorithms yield substantially better performance than ZeroR and random guess. Abalone conservation often focuses on protecting the “Small numbers of young abalones” left from overfishing [7]. In that case, the algorithms may provide useful conservation tools as their precision-recall performance for the infant class is strong at ≥ 68.7%. In particular, the IBk model may be used since it gave the best precision of 73.3% for the infant class.

B. What I learned from this process

This prediction exercise helped me further appreciate the point made in Siegel’s *Predictive Analytics* book that “Data is always predictive” [8, Ch. 3, p. 79] and that “A little prediction goes a long way.” [8, Ch. 1, p. 35] Given the choice of a random guess with approximately 33.3% accuracy, the models increased overall accuracy substantially to >50%. In fact, for infant abalones, the accuracy is even higher with precision reaching 73.3% in the IBk model.

It seems evident from this process that we can always find a model that will do better than mere guessing. Prediction models provide useful tools to increase efficiency and productivity all at the readily attainable cost of utilizing an open source toolkit such as Weka.

C. The process has deepened my knowledge of the course material

Going through this exercise provided me with hands-on experience on the nuts and bolts of a prediction task. It provided me with a real world application of the concepts and algorithms discussed in class from sourcing data, asking the *right* question about the data, preprocessing the data, choosing suitable learning algorithms for the data to interpreting results. I now have more confidence and practical knowledge to approach any dataset and attempt to make meaning out of it.

Acknowledgment

I would like to thank Professor Piotr M. Szczurek for his suggestion to predict the sex of abalones. I would also like to thank the UCI repository for the Abalone Dataset.

References

[1] Blake, C., Keogh, E. and Merz, C. J., UCI Repository of Machine Learning Databases, Irvine, CA: University of California, Department of Information and Computer Science, Irvine, CA, 1998. Available: https://archive.ics.uci.edu/ml/datasets.html. [Accessed: Mar. 10, 2015].

[2] Iqbal, Naveed Hussein, and Georgios C. Anagnostopoulos. “Multinomial Squared Direction Cosines Regression.” Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011.

[3] R. Remco, Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse, “WEKA Manual for Version 3-7-3,” The University of Waikato, New Zealand, 2010.

[4] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. “Data Mining: An Overview from a Database Perspective,” IEEE Transactions on Knowledge and Data Engineering, 8(6):866-883, December 1996.

[5] Faloutsos, C., Korn, F., Labrinidis, A., Kotidis, Y., Kaplunovich, and A. Perkovic, D., “Quantifiable Data Mining Using Principal Component Analysis,“ Institute for Systems Research, University of Maryland, College Park, MD, Technical Report CS-TR-3754, 1997.

[6] S. Inamdar, S. Narangale, and G. Shinde, “Preprocessor agent approach to knowledge discovery using Zero-R algorithm,” International Journal, vol. 2, 2011.

[7] California Department of Fish and Wildlife, “California Abalone Information,” *California Department of Fish and Wildlife*. Available: http://www.dfg.ca.gov/marine/invertebrate/ab_info.asp. [Accessed: Mar. 4, 2015].

[8] Siegel, E., *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, NJ: John Wiley & Sons, Inc., 2013.