Abalone Age Prediction Using Machine Learning

Chapter *in* Communications in Computer and Information Science · April 2022

DOI: 10.1007/978-3-031-04112-9_25

4 authors, including:

Alaa Ali Hameed
Istinye University
123 PUBLICATIONS **976** CITATIONS

SEE PROFILE

Akhtar Jamil
National University of Computer and Emerging Sciences
127 PUBLICATIONS **1,391** CITATIONS

SEE PROFILE

Abalone Age Prediction Using Machine Learning

Seda Guney[1], Irem Kilinc[1], Alaa Ali Hameed[1] and Akhtar Jamil[2]

[1] Istanbul Sabahattin Zaim University, Istanbul 34303, Turkey
seda.guney@std.izu.edu.tr, irem.kilinc@std.izu.edu.tr,
alaa.hameed@izu.edu.tr
[2] National University of Computer and Emerging Sciences, Islamabad, Pakistan
akhtar.jamil@nu.edu.pk

Abstract. Abalone is a marine snail found in the cold coastal regions. Age is a vital characteristic that is used to determine its worth. Currently, the only viable solution to determine the age of abalone is through very detailed steps in a laboratory. This paper exploits various machine learning models for determining its age. A comprehensive analysis of various machine learning algorithms for abalone age prediction is performed which include, backpropagation feed-forward neural network (BPFFNN), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, Gauss Naive Bayes, and Support Vector Machine (SVM). In addition, five different optimizers were also tested with BPFFNN to evaluate their effect on its performance. Comprehensive experiments were performed using our data set.

Keywords: Machine Learning, Neural Networks, Abalone, Back Propagation Neural Networks

1 Introduction

Abalones are types of single-shelled marine snails found in the cold coastal waters worldwide, majorly found along the coastal regions of some countries such as Australia, Western North America, South Africa, New Zealand, and Japan [1]. The age of the abalone is highly correlated to its prices as it is the sole factor used to determine its worth [2]. However, determining the age of abalone is a highly involved process that is usually carried out in a laboratory.

Technically, rings are formed in the inner shell of the abalone as it grows gradually at a rate of one ring per year. To get access to the inner rings of an abalone, the shell's outer rings need to be cut. After polishing and staining, a lab technician examines a shell sample under a microscope and counts the rings. Because some rings are hard to make out using this method, 1.5 is traditionally added to the ring count as a reasonable approximation of the age of the abalone [1]. Knowing the correct price of the abalone is important to both the farmers and consumers. In addition, knowing the correct age is also crucial to environmentalists who seek to protect this endangered species.

Due to the inherent inaccuracy in the manual method of counting the rings and thus calculating the age, researchers have tried to employ physical characteristics of the abalone such as sex, weight, height, and length to determine its age. Thus, by applying

machine learning on a dataset containing a large number of training samples of physical measurements of abalone, its age can be predicted quickly and more accurately.

Machine learning algorithms are data-driven approaches that can effectively recognize certain patterns. Over the last decade, machine learning techniques have been successfully applied across various domains such as for Unicode symbols identification [3], classification of large data sets [4], ordinal classification [5], etc. Among machine learning approaches, the most successful and widely used techniques include Artificial neural networks (ANN), KNN, random forest, Gauss Naïve Bayes and SVM. These techniques have been widely used to solve many pattern recognition problems, which include both classification and regression.

The literature review indicates that the artificial intelligence approaches for abalone age prediction are not sufficient. A detailed search for literature for abalone on popular databases such as Google Scholar, Web of Science, and Scopus showed that only a few papers were found. In [6] authors proposed a feed-forward multi-layer perceptron model to predict the age of abalone. The model was trained with the Lavenberg-Marquardt backpropagation algorithm. Experimental demonstrated that the error rate gradually decreased as the number of hidden layers increased. Obtained results showed the robustness and effectiveness of this model as it achieved high classification accuracy.

Similarly, in [7], the authors proposed a regression-based artificial neural network to determine abalone age. They employed a relatively shallower network with three hidden layers for prediction. The physical measurements of the abalone were used as features, and the results were highly accurate.

Similarly, in [8], the authors applied a convolutional neural network (CNN) model. The authors experimented with various network architectures and different types of convolutions. Also, the deep learning-based approach was compared with conventional machine learning methods. The deep learning-based method achieved 79.09% classification accuracy, which was much better than conventional approaches.

In this study, we exploit the power of seven different machine learning algorithms for abalone age prediction. The algorithms included BPFFNN, K-nearest neighbor, Naive Bayes, decision tree, random forest, Gauss naive Bayes, and support vector machine. In addition, five different optimizers were also tested with BPFFNN to evaluate their effect on its performance.

The rest of the paper is organized as follows: Section II describes the related works. Section 3 presents the proposed method and the dataset used. In Section 4, obtained experimental results are presented. Finally, we complete the paper with concluding remarks in section 5.

2 Materials and Methodology

2.1 Dataset

The abalone dataset was obtained from UCI Machine Learning Repository [9]. The dataset consists of 4176 samples with eight numerical features along with labels. These features represent the physical characteristics of abalone, such as gender, length, height,

diameter, whole weight, shucked weight, viscera weight, shell weight, and rings. The distances were measured millimetres, while the weight values were measured in grams. Gender, a categorical feature, has been converted into numerical features, and all values were transformed to numerical.

We categorized the abalone into three age groups: G1: below 7, G2: between 7 and 16, G3: above 16. These ages groups were represented as G1=1, G2=2 and G3 = 3. For some algorithms, such as ANN the input labels were proceeded to represent them as one hot-vector representation.

2.2 Classification

This section briefly summarizes the classification and optimization algorithms used in this study.

Artificial Neural Networks (ANNs) ANNs consist of neurons arranged in layers. The first layer is called the input, while the last layer is referred to as the output layer. Hidden layers are placed between input and output layers to map the input to the output. The perceptron equation is as follows:

$$y = \varphi\left(\sum_{i=1}^{n} W_i * x_i\right) \tag{1}$$

Where φ is the activation function, X is the input, and W is the weight vector.

$$p_c = \frac{eW^{r+b}}{\sum_{i=1}^{L} e\, W_i{}^{r+b_i}} \tag{2}$$

We exploited both categorical hinge (3) as well as mean squared error (4) as loss functions in our experiments. In order to find an optimal accuracy and evaluate the effectiveness of optimization algorithms, proposed neural network model were trained separately with Stochastic Gradient Descent, Adagrad, RMSprop, Adam, Nadam optimization algorithm respectively.

$$\text{MSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

In this study, five layers were used; one input, three hidden, and one output layer. The number of neurons on input and output were selected according to input feature size and output classes. The number of neurons at the hidden layers were empirically calculated and was set to 70. ReLu activation function was used in all layers except the output layer, which used softmax. The learning rate was set to 0.01.

K-Nearest Neighbor KNN is one of the oldest, simple and efficient supervised machine learning algorithms. It has been applied in various applications, including problems of classification or regression. It is a non-parametric classification method that means it does not require any prior information about the data distribution.

KNN employs distance measures to learn and classify the samples. We applied three distance measurements Euclidian, Minkowski , and Manhattan. We found that Euclidian distance was better for our data set [10]. For KNN algorithm, the selection of K, which

is the number of neighbours, plays a crucial role. For any point p, the closest K neighbours are selected, and the decision is based on the majority vote of these k-neighbours.

Naive Bayes Naïve Bayes is a classical classification algorithm that is based on Baye's Theorem. It assumes that the predictors are independent. The main objective is to maximize the posterior probabilities for each class. Theoretically, it results in a minimum error rate and has resulted in worked well for various real-world applications.

Consider the input data X with hypothesis H, prior P(H|X); the posterior probability can be estimated as follows

$$P(H|X) = P(X|H)P(H)/P(X) \tag{4}$$

Decision Tree A decision tree is a simple supervised learning algorithm that can be employed for both classification and regression tasks. It continuously split the data into smaller subset based on some criteria. Then a voting mechanism is followed to make the final decision [11] [12]. There are two main types of decision trees: classification trees and regression trees. The classification trees are the ones where the output variable is discrete, while in the case of regression trees, the output variable is continuous.

To construct the decision tree, entropy and information gain are generally employed [13],[14]. The process iteratively continues splitting the data at each node until the leaves are pure. To avoid the overfitting problem, a limit on the depth of the decision tree is also introduced.

The information gain for each attribute is calculated using the following equations:

$$Gini\ index = 1 - \sum_{i=1}^{c} {p_i}^2 \tag{5}$$

$$Gain\ ratio = \frac{Information\ gain}{Split\ information} \tag{6}$$

Random Forest Random forests are ensemble techniques that are widely used for both classification and regression. They employed multiple decision trees for training and testing. This makes them more robust compared to a single decision tree. The random forest algorithm generates an independent random vector θ_k from the previous random vectors which is then distributed to all trees. The trees are grown during training stage using the random vector θ_k, that will result in a tree-structured classifiers $\{h(x, \theta_k), k = 1, \ \ldots\}$. The generalization error is calculated by [16]:

$$\mathrm{PE} * = \ \mathrm{PX,Y}(\mathrm{mg(X,Y)} \ < \ 0) \tag{7}$$

where X and Y are random vectors and mg is the margin function.

Gauss Naive Bayes The Gaussian Naïve Bayes classifier is a special case of Naïve Bayes in continuous case.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{8}$$

In practice, we assume that each of the probability function has a Gaussian distribution, only need to calculate mean μ and variance σ^2 to obtain the density in Eq. (10).

$$(x_i|\omega_i) = f(x_k) \tag{9}$$

This classifier is called Gaussian Naïve Bayes. In which we need to compute the mean μ_i and variance σ^2 of each training samples of classes ω_i.

Support Vector Machines. Support Vector Machine (SVM) is a non-paramatric classifier. Basically, SVM is basically a linear binary classifier that assigns a class label to test data for classification. It tries to maximize the margin between the two classes. However, it can also be adopted to non-linear data by applying kernel techniques. An SVM constructs a hyperplane in an infinite-dimensional space which is used for classification or regression. This hyperplane is obtained by finding the maximum separation between two classes with the nearest training data points. These data points are termed as support vectors. SVM is a highly powerful model that can produce optimal accuracy even in the presence of lower training samples.

3 Experimental Results

A number of experiments were performed to obtain the optimal parameters for each model. All the experiments were performed on the standard Intel (R) Core (TM) İ5-7200U CPU @ 2.50GHz computer in an Anaconda environment with Python as the programming language. The training dataset consists of 4176 samples. These samples were divided into training consisting of 2923 samples (70%) and testing 1253 samples (30%) subsets.

The Training accuracy with different optimizers is shown fig. 1. BPFFNN model obtained high accuracy for both training and testing. Moreover, Adadelta optimizer scored better compared to other optimizers with BPFFNN (89% training and 88% testing). The figure shows that all optimizers produced similar results except Sgd optimizer.

In fig. 2 the convergence of five different optimization algorithms is illustrated in terms of training loss over the epochs. BPFFNN model had a lower training loss with Adagrad optimizer. Sgd starts with a rapid descent, but after 150 epoch stops improving. Rmsprop, Adadelta and Adam optimizers seem to perform almost the same.

Table 1 shows the confusion matrix for a multiclass classification problem with three classes (1, 2 and 3). As seen in the table, TP_1 is the number of true positive samples in the class 1, that is, the number of samples that are correctly classified from class 1. E_{12} is misclassified samples, i.e., the samples from class 1 that were incorrectly classified as class 2. Accordingly, the false negative in the 1 class (FN_1) is the sum of all class 1 samples that were incorrectly classified as class 2 or 3, i.e., is the sum of E_{12} and E_{13}.

Briefly, FN of any class is equal to the sum of a row except value TP. The false positive (FP) of any class is equal to the sum of a column except the value TP. The true negative (TN) of any class is equal to the sum of values except row of true class and the column of predicted class.

$$\mathrm{FN}_{1=}\mathrm{E}_{21} + \mathrm{E}_{31} \tag{10}$$

$$\mathrm{FP}_1 = \mathrm{E}_{12} + \mathrm{E}_{13} \tag{11}$$

$$TN_1 = TP_2 + E_{32} + E_{23} + TP_3 \tag{12}$$

Table 2 summarizes the accuracy of all the classifiers. Generally, all classifiers performed equally well, except the Gauss Naive Bayes, which obtained relatively lesser accuracy (60.88%). Moreover, the Random Forest classifier produced the highest performance on our dataset (87%) followed by SVM, which achieved an accuracy of 86.76 %. Furthermore, KNN and Decision tree classifiers reach almost equal accuracy 86.28%, 86.44%, respectively. Compared with the other classifiers, the performance of the proposed model was relatively better. From the obtained results, we can conclude that the BPFFNN reached the best accuracy in the abalone age prediction task.

Table 1 The confusion matrix for a multiclass classification problem with three classes.

Pred / True	G1	G2	G3
G1	TP1	E21	E31
G2	E12	TP2	E32
G3	E13	E23	TP3

G1: age < 7, G2: 7 ≤ age ≤, and G3: age > 16

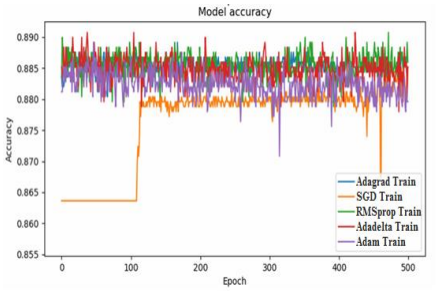

Fig.1. Training accuracy with different optimizers

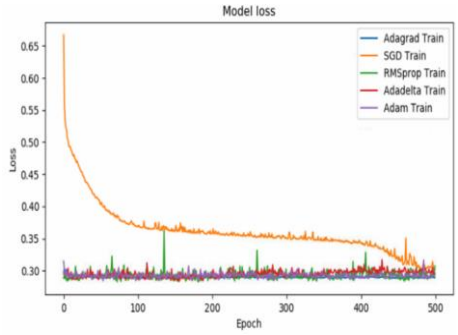

Fig. 2. Training loss for each optimizer.

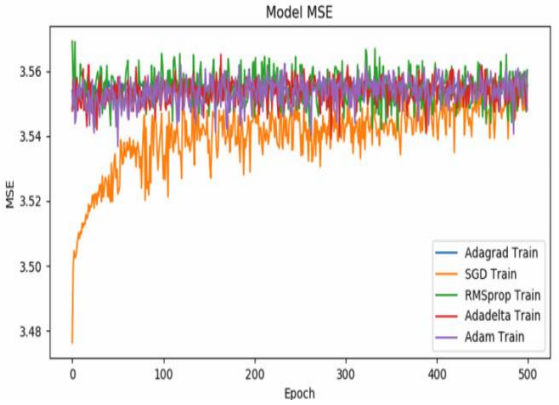

Fig. 3 Mean Square Error for different optimizers

Table 2 Comparison of applied algorithms.

No	Method	Accuracy %
1	KNN	86.28
2	Gaus Naive Bayes	60.84
3	Decision Tree	86.44
4	Random Forest	87.00
5	Support Vector Machine	86.76
6	**Proposed (BPFFNN)**	**88.25**

The confusion matrix of the Random Forest algorithm is presented in Table 3. Obtained results demonstrate that it performs the best results for Group-2 class of the dataset. There only 8 data in Group-2 that are misclassified. One thousand forty-one data in Group 2 are classified correctly. Only 3 data classified correctly in Group-3, which shows it does not perform well.

Table 3 Confusion matrix for Random Forest.

Pred / True	G1	G2	G3
G1	35	31	0
G2	7	1041	1
G3	0	136	3

G1: age < 7, G2: 7 ≤ age ≤, and G3: age > 16

Confusion matrix of SVM algorithm is presented in Table 4. Obtained results demonstrates that, the SVM algorithm gave the worst results after the Gauss Naïve Bayes algorithm for class 2 (between 7 and 16 age of abalone). 59 data in class 2 are misclassified, that is, FN_2.

Table 4 Confusion matrix for SVM.

Pred / True	G1	G2	G3
G1	44	22	0
G2	11	990	48
G3	0	85	54

G1: age < 7, G2: 7 ≤ age ≤, and G3: age > 16

Confusion matrix of KNN algorithm is presented in Table 5. Obtained results demonstrates that, 1006 data in class 2 are classified correctly, i.e. TP_2. Only 35 data

classified correctly in the 3(above 16 age of abalone) class, that is, TP_3. The KNN and decision tree algorithms gave the worst results for class 1.

Table 5 Confusion matrix for KNN.

Pred / True	G1	G2	G3
G1	41	25	0
G2	11	1006	32
G3	0	104	35

G1: age < 7, G2: 7 ≤ age ≤, and G3: age > 16

Confusion matrix of Gauss Naive Bayes algorithm is presented in Table 6. Obtained results demonstrates that, relevant algorithm preform the best results for 1(below 7 age of abalone) and 3 (above 16 age of abalone) class of the dataset. There only 4 data in class 1 (below 7 age of abalone) that are misclassified, that is, FN_1. The Gauss Naive Bayes algorithm gave the worst results for class 2. Only 629 data classified correctly in the 2((between 7 and 16 age of abalone)) class, i.e. TP_2. 420 data in class 2 are misclassified, that is, FN_2.

Table 6 Classification result for Gauss Naive Bayes

Pred / True	G1	G2	G3
G1	62	4	0
G2	108	629	312
G3	0	67	72

G1: age < 7, G2: 7 ≤ age ≤, and G3: age > 16

The confusion matrix of the decision tree algorithm is presented in Table 7. Obtained results demonstrate that 1018 data in class 2 are classified correctly, i.e. TP_2. Only 31 data in class 2(between 7 and 16 age of abalone) are misclassified, that is, FN_2. The decision tree algorithm gave the best results after the random forest algorithm for class 2. Only 24 data classified correctly in the Group 3. The decision tree algorithm gave the worst results after the random forest algorithm for class 3. While for 3 class, it does not perform well.

Table 7 Classification result for Decision Tree.

Pred / True	G1	G2	G3
G1	41	25	0
G2	10	1018	21
G3	0	115	24

G1: age < 7, G2: 7 ≤ age ≤, and G3: age > 16

The overall results obtained for abalone classification using the six conventional classifiers were satisfactory except Gauss Naive Bayes classifier. The proposed BPFFNN outperformed all other classifiers in terms of classification accuracy. In addition, we compared our approach with CNN based method proposed by authors in [8], which reported 79.09% accuracy. We believe that for simple datasets such as the one we used in this study, the conventional machine learning approaches are more effective than deep learning-based approaches. Even though deep learning-based approaches have shown high classification accuracy for many problems, yet they are data intensive. We prefer conventional machine learning approaches over deep learning methods for both ease of implementation and classification accuracy in scenarios like this where the dataset is small.

4 Conclusion

This paper focused on abalone age prediction using machine learning techniques. Six state-of-the-art models were employed, which are commonly used for classification tasks. We further quantified the performance of BPFFNN using five different optimizers to select the one that performs best. For our dataset, BPNN yielded the highest accuracy (88%), followed by the random forest classifier (87%). In the future, we would like to extend our method by employing an automatic feature extraction step from images of the abalone for its age prediction.

References

[1] Abalone: https://en.wikipedia.org/wiki/Abalone

[2] Hossain, M, & Chowdhury, N Econometric Ways to Estimate the Age and Price of Abalone. Department of Economics, University of Nevada (2019).

[3] UCI Machine Learning Repository, Abalone dataset: https://archive.ics.uci.edu/ml/datasets/Abalone

[4] A.B.Karthick Anand Babu, Design and Development of Artificial Neural Network Based Tamil Unicode Symbols Identification System. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, (2012).

[5] Alsabti, K., Ranka, S., & Singh, V CLOUDS: A decision tree classifier for large datasets (1999).

[6] K. Jabeen ve K. Ahamed, "Abalone Age Prediction using Artificial Neural Network," IOSR Journal of Computer Engineering, vol. 18, no. 05, pp. 34–38, (2016).

[7] Misman, M. F., Samah, A. A., Ab Aziz, N. A., Majid, H. A., Shah, Z. A., Hashim, H., & Harun, M. F. (2019, September). Prediction of Abalone Age Using Regression-Based Neural Network. In 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 23-28). IEEE.

[8] Sahin, E., Saul, C. J., Ozsarfati, E., & Yilmaz, A. Abalone Life Phase Classification with Deep Learning. In 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 163-167 (2018).

[9] UCI Machine Learning Repository, Abalone dataset: https://archive.ics.uci.edu/ml/datasets/Abalone

[10] Bhatia, N. Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085 (2010).

[11] Kerber R ChiMerge: discretization of numeric attributes. In: Proceedings of the tenth national conference on artificial intelligence (1992).

[12] Han, J., and M. Kamber. "Data Mining: Concepts and Techniques, 550." (2000).

[13] Bramer, Max. Principles of data mining. Vol. 180. London: Springer, (2007).

[14] Leung, KwongSak, et al. "Data mining on dna sequences of hepatitis b virus." IEEE/ACM transactions on computational biology and bioinformatics 428-440 (2009).

[15] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." 2008 IEEE/ACS international conference on computer systems and applications. IEEE, 2008.

[16] L. Breiman, Random forests, Mach. Learn. 45 45 -49 (2001).