Age of Abalones using Physical Characteristics: A Classification Problem

ECE 539 Fall 2010 Project Report

Hiran Mayukh

Department of Electrical and Computer Engineering
University of Wisconsin-Madison

mayukh@wisc.edu

Abstract

Abalones [1], also called ear-shells or sea ears, are sea snails (marine gastropod mollusks) found world-wide. The age of an abalone can be determined by counting the number of layers in its shell. However, age determination is a cumbersome process: It involves cutting a sample of the shell, staining it, and counting the number of rings through a microscope. A data set provided by the University of California Irvine Machine Learning Repository [2] consists of physical characteristics of abalones and their ages. This study is a classification problem that aims to predict the ***age range*** *of abalones using their physical characteristics.*

The data set is pre-processed to transform the problem of predicting the age to a classification problem. Two clustering algorithms are used to cluster the training data set without supervision. Cluster centers are then grouped together and assigned class labels based on votes by data points within each cluster region. The testing data set is used to estimate accuracy of the classifier. Experiments are also run to obtain an order of physical characteristics reflecting their contribution to classification accuracy.

1. Introduction

Machine learning (ML) algorithms are used to recognize patterns and make decisions based on empirical data. The problem of classification of a data set, that is, assigning a class label to each sample of the data set can be complex, especially if the data set large or if the data set has a large feature dimension. If the feature vector representing a sample of data is of *n* dimensions, the problem of classification boils down to carving out regions in *n*-dimensional feature space with the understanding that any point within a space is to be assigned a certain class label. In this study, unsupervised clustering algorithms are used to divide feature space into a large number of regions, and based on the class labels of training set samples present in each region, each region (specified by the region's clustering center) is assigned a class label.

In classification problems with feature vector dimensions more than one, an interesting question is the relative importance of each feature to classification. However, features may be correlated with each other in complex ways, and the contribution of each feature to enabling better classification may be difficult to determine. This study experiments with reduced number of features (i.e., by elision of different combinations of features) in order to generate an approximate ordering of data set features used in the classification problem dealt with in the paper.

This paper is organized as follows: Section 2 introduces the data set and the preprocessing done in order to recast it as a classification problem. Section 3 describes the methodology, section 4 contains the experiments and results, and section 5 concludes the paper.

2. Data Set

The University of California, Irvine Machine Learning Repository [2] provides a data set consisting of 4177 samples of physical characteristics of abalones and their age. Abalones are sea-snails that are fished for their shells and meat. Scientific studies on abalones require knowing the age of an abalone, but the process of determining age is complicated. It involves measuring the number of layers of shell (“rings”) that make up the abalone’s shell. This is done by taking a sample of shell, staining it and counting the number of rings under the microscope. To circumvent the cumbersome process, this data set provided has been used to build learning algorithms to predict age using easily and quickly measurable physical characteristics. This project uses this data set recast as a classification problem, rather than a prediction problem.

The data set consists of 8 features and the number of rings (which is directly related to the age). The 8 features are sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight. For ease, we refer to these as F1, F2 … and F8 in this paper.

The number of rings varies from 1 to 29. This could be looked at as a class label that can take 29 possible values. To reduce the time taken by code to run experiments, the number of classes is reduced from 29 to 8. Data samples with number of rings from 1 to 4 are assigned class label 1, 5 to 8 are assigned class label 2, and so on, until 24 to 28 gets class label 7 and samples with 29 rings is said to be of class 8. Each class, thus, forms an age range. Figure 1 shows the histogram of the entire data set.

Other minor preprocessing done to the data set include converting the three possible values of feature sex (male, female, infant) to numbers; and dividing the data set into a training set and a testing set on the same lines as other studies with this data set [4,5]. The first 75% of samples (3133) form the training set and the remaining (1044) form the testing set.

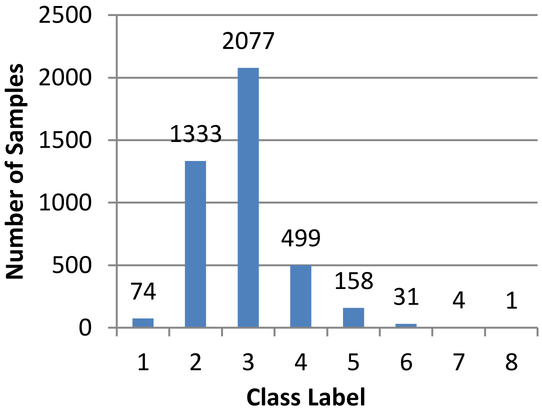

Figure 1. Histogram of the Abalone data set

3. Methodology

Two clustering algorithms, the k-means algorithm and a hierarchical clustering algorithm are developed in MATLAB [6], and run to generate a large number of cluster centers. All test data samples that lie within Voronoi regions created by the cluster centers are then polled to find the most common class label, and the cluster center (and hence the cluster’s Voronoi region) is assigned that class label. The idea is that if the number of clusters is made large enough, we get a fine grained division of the feature space based on class label, and hence can be used to predict the class label of testing data samples based on the region it lies in. The algorithm is spelt out in Table 1:

The experiment has to be repeated many times because the clustering algorithms are not deterministic – the clustering configuration depends strongly on initial positions of cluster centers, which are randomly initialized to decrease the chances of repeatedly falling into a local minimum cluster center configuration. This is also why the highest accuracy achieved (and not the average) is taken. It is interesting to note that when run many times (>50) with random initial state, the highest accuracy obtained occurs multiple times, a sign that the accuracy reflects that of the globally best classification that the algorithms can provide.

```
1 Read in the data set
2 Process age to convert to class
  labels
3 Split into training and testing
  data
4 Use a classification algorithm
  to classify training data into
  M regions (M >> 8)
5 Poll the region around each
  cluster center for the class of
  the closest point, and build a
  cluster to class label mapping
  for all clusters
6 For each sample in the testing
  data, find the closest cluster
  center, then find the
  corresponding class label using
  the mapping obtained from the
  previous step. Assign the
  testing sample that class label
7 Calculate accuracy by comparing
  with actual class label of the
  testing samples
8 Repeat steps 4-8 a large number
  of times and pick the highest
  classification rate achieved
```

Table 1. Steps in the generation of classification accuracies

The hierarchical clustering algorithm used is a variation on the k-means algorithm that starts out with number of clusters equal to 2. After a classification decision is made, the larger of the two clusters (with larger number of data points inside its region) is split. This process is continued iteratively until the number of clusters reaches M, which is much larger than the number of class labels (we chose a value of 80, which is an order of magnitude larger than the number of classes).

4. Experiments and Results

4.1. Accuracy Results

The classification accuracies achieved on the testing data set and the time taken for classification by the two approaches are shown in Table 2.

The accuracy of the hierarchical clustering algorithm is *far lower* than that of k-means. We think this is due to the implementation of k-means on MATLAB, which can generate cluster regions with no points inside them – this behavior seems to happen more often when the argument specifying number of clusters passed to the MATLAB function *kmeans* is low. For hierarchical clustering, this argument is always 2. Many regions created by our hierarchical clustering algorithm are empty – so even if number of cluster regions M is identical in the two clustering approaches, the actual number of cluster regions with points in them (a requirement to be able to map the cluster region to a class label) is lower in the hierarchical approach. This is the cause for the large difference in accuracies.

Clustering Algorithm	**Classification Accuracy (%)**	**Time (s)**
k-means	61.78	4.85
Hierarchical	6.23	6.34

Table 2. Classification accuracies and run times in seconds

The remaining experiments in this project use only the k-means clustering algorithm.

4.2. Feature Vector Results

To rank the features in a partial order of their importance to predicting the age range, we first run the classification experiments after preprocessing the data set to exclude one feature at a time. The x-axis in Figure 2 is the feature that has been elided from the data set.

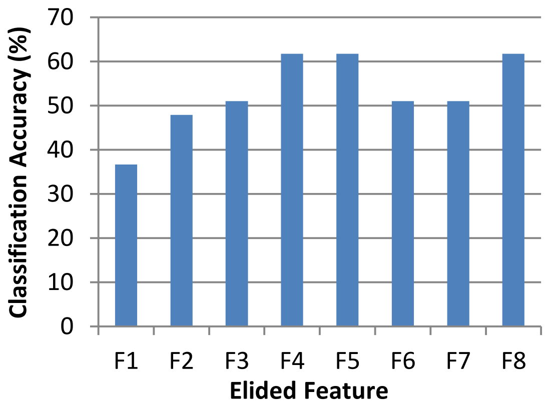

Figure 2. Classification accuracy plotted against feature elided from data set

Features F1 and F2 are the first and second most important feature for age classification, but a total ordering of features cannot be deduced from this graph – F4, F5 and F8 degrade accuracy by the same amount, as do F3, F6 and F7.

To break this grouping, experiments are run leaving 2 out of 3 at a time among F4, F5 and F8. Figure 3 plots this experiment’s classification accuracies; the x-axis shows the two features that have been removed. Note that F1 and F2 are still part of the data set.

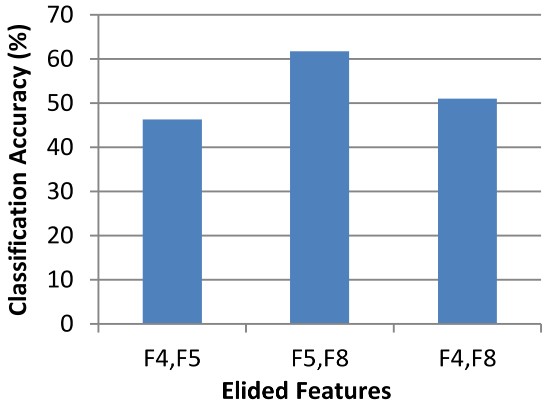

Figure 3. Classification accuracy plotted against features elided from data set

Now it can be seen that among F4, F5 and F8, the ordering of importance of features are F4 > F5 > F8. A similar experiment is run for F3, F6 and F7, and the results are shown in Figure 4.

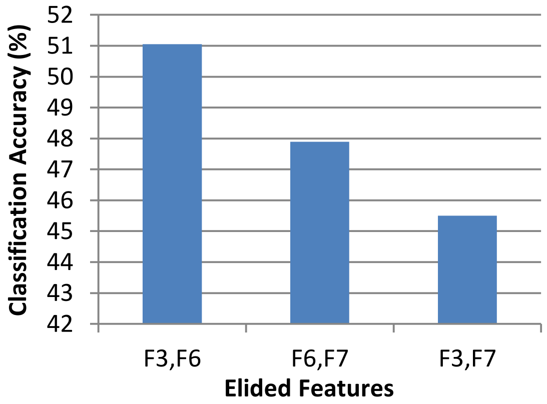

Figure 4. Classification accuracy plotted against features elided from data set

Thus, the ordering of the features with respect to importance is F1, F2, F4, F5, F8, F7, F3 and F6 (Table 3).

F1	Sex
F2	Length
F4	Height
F5	Whole weight
F8	Shell weight
F7	Viscera weight
F3	Diameter
F6	Shucked weight

Table 3. Features in the Abalone data set in order of importance in predicting the age range

5. Conclusion

This project studied classification through unsupervised clustering algorithms by applying them to predict the range of age of Abalones. Experiments were also run and interpreted to rank the features in how much they reflect on the age of the abalone.

References

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

[2] UCI Machine Learning Repository: Abalone Data Set. http://archive.ics.uci.edu/ml/datasets/Abalone

[3] Abalone Data Set README file. http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names

[4] Extending and benchmarking Cascade-Correlation. Sam Waugh (1995), PhD thesis, Computer Science Department, University of Tasmania.

[5] A Quantitative Comparison of Dystal and Back-propagation. David Clark, Zoltan Schreter, Anthony Adams. Australian Conference on Neural Networks (ACNN'96)

[6] MATLAB. www.mathworks.com/products/matlab