University Of Southampton

Department of Electronics and Computer Science

COMP 2039

Artificial Intelligence

Fine Tuna

A report on the implementation and accuracy of a KNN/Bayes Naïve classifier to determine the activity of a tuna over a 12 hour period.

Author:

Ben Kneebone

bk8g11

Date of submission: 17/04/13

Contents

Data Standardisation

As the range of each dimension differs, raw data must be standardised so that each dimension shares the same range so a Euclidean Distance Metric can be applied without giving any dimension more importance than any other dimension.

Method used to standardise each piece of raw data in the test set:

$$\boxed{Standardised\ value = \frac{Unstandardised\ value - Mean\ value\ of\ dimension}{Standard\ Deviation\ of\ dimension}}$$

Equation 1 – Standard Score

The various mathematical functions necessary are performed by the class ArrayListMath.

The KNN Classifier

Incrementally built with the primary milestones:

1) KNN-1, 8 dimensions

- The difference between the new point and each point in the standardised test set is summed across the 8 dimensions.
- Distance compared to a double variable holding the distance of the nearest point in the test set to the new point found so far.
    - If this distance is lower, the double and a String value representing classification is updated to that of the new nearest test set point.
- This String is returned when all of the test set points have been checked.

2) KNN-I, 8 dimensions

- Array stores i nearest points as instances of DistClass (Stores distance and classification -See Appendix 1).
- Difference between sample and each standardised test set point summed across 8 dimensions.
- Check if one of the nearest i points (Compare to Array values).
- Classification returned depends on polling methods used (p5).

3) KNN-i classifier using different dimensions

- Same as 2, except only summed across certain dimensions (Those set to true in the boolean[] parameter of classify()).

4) Bayes Naïve classifier

- The steps in 2/3) are executed, and Bayes Naïve was added as a polling option.

Polling Methods

Classification methods take an int 1-4 indication which polling method should be used.

	Method
1	$$\sum_{x=0}^{i}(Tally\ of\ classification\ of\ nearest\ neighbour\ x) + 1$$ Equation 2 – Normal Polling
2	$$\sum_{x=0}^{i}(Tally\ of\ classification\ of\ nearest\ neighbour\ x) + (i - x)$$ Equation 3 – Weighted Polling
3	$$\sum_{x=0}^{i}(Tally\ of\ classification\ of\ nearest\ neighbour\ x) + (i - x)^2$$ Equation 4 – Weight-Squared Polling
4	Bayes Naïve. • Frequency of each classification in the test set tallied on classifier creation (Maintained by one left out methods – Incrementing/Decrementing as necessary). Method one with a final step: $$Probability\ of\ Classification = \frac{Frequency\ of\ classification\ in\ nearest\ i\ neighbours}{Frequency\ of\ classification\ in\ test\ set}$$ Equation 5 – Bayes Naïve Classifier

Figure 1 – Polling Methods

- Highest votes/probability is declared the winner.
- In a draw:
  - Deduct votes/probability given to the furthest of the I neighbours.
  - Check for winner. Repeat deduction if necessary for next furthest neighbour. Guaranteed a winner at the nearest neighbour if necessary.

Testing

Testing is undertaken using the one left out method. It is performed on a unique I, method and dimension subset combination. This means trying to classify each of the 76 samples we are given using the other 75:

- Remove one of the data points (Store as sample).
- Find the expert classification given to that point.
- Standardise the test set and sample (As a sample has been removed so Mean/Standard Deviation have changed).
- Classify the sample we have taken.
- Check if the classification matches with the expert prediction.
- Return a percentage of correct classifications.

The harness methods findBestAccuracy() and findAverageAccuracy run a number of one left out tests.

findBestAccuracy(int iMin, int iMax, int method)

One left out test for all dimension subsets for every I between and including iMin and iMax using the specified method. Uses a double variable to store the highest percentage so far, an int variable to store the I that produced it and an index that maps to a dimension subset. These are then printed out upon completion to show the best configuration and the percentage produced.

findAverageAccuracy(int I, int method)

One left out test for all dimension subsets for I using the specified method. Stores the returned accuracies in an ArrayList<Double>. The mean of the percentages is returned upon completion.

Results

KNN-1 Classifier all dimensions

Method	
1	69.74
2	69.74
3	69.74
4	69.74

Table 1 – KNN-1 All Dimension results

Comments:

- Identical as all return the classification of the nearest neighbour.

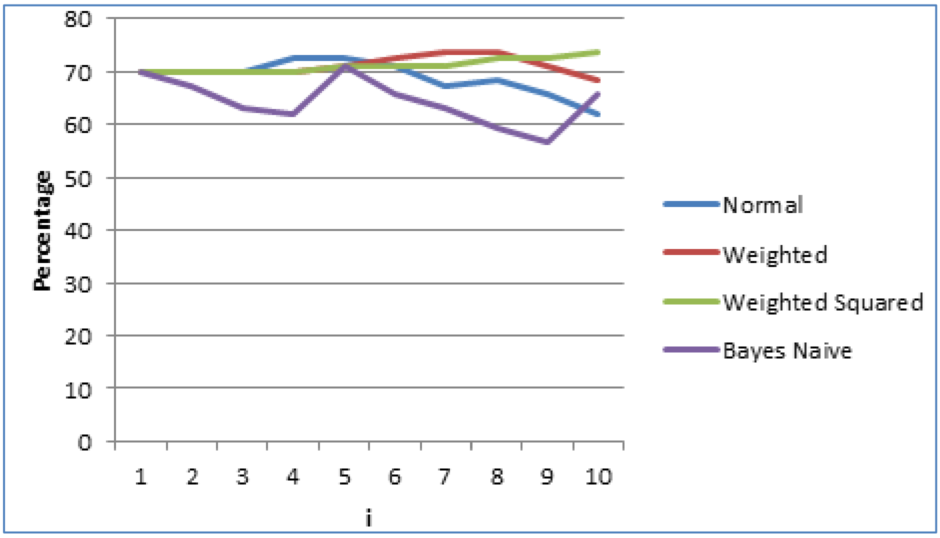
KNN-I classifier all dimensions

i	Method 1	Method 2	Method 3	Method 4
1	69.74	69.74	69.74	69.74
2	69.74	69.74	69.74	67.11
3	69.74	69.74	69.74	63.16
4	72.37	69.74	69.74	61.84
5	72.37	71.05	71.05	71.05
6	71.05	72.37	71.05	65.79
7	67.11	73.68	71.05	63.16
8	68.42	73.68	72.37	59.21
9	65.79	71.05	72.37	56.58
10	61.84	68.42	73.68	65.79
Average:	**68.82**	**70.92**	**71.05**	**64.34**

Table 2 – KNN-I all Dimensions results

Comments:

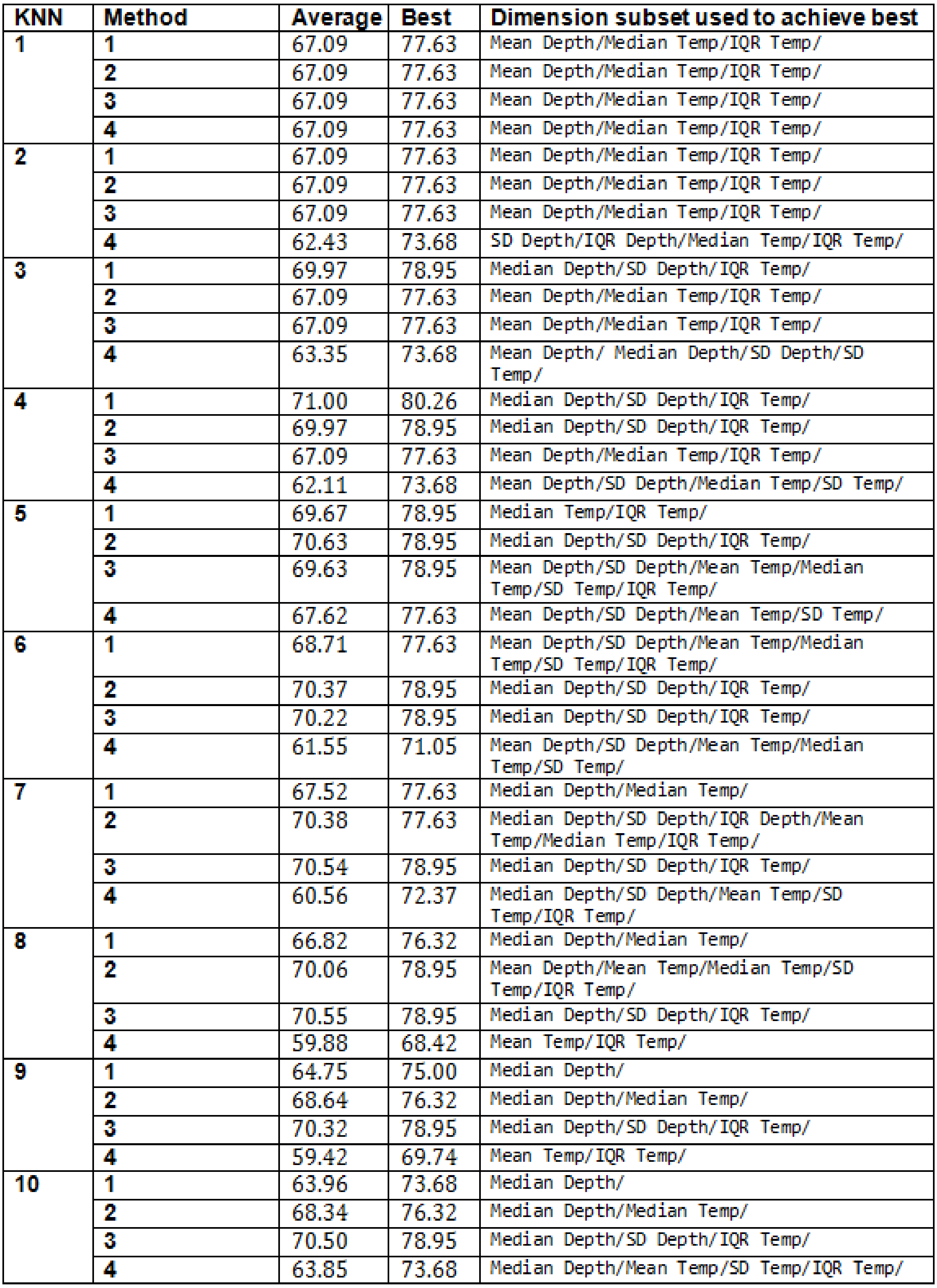
- Bayes Naïve often the worst result, put down to similar numbers of each classification in the test set.
- Weighted squared becomes better with increasing i. Nearest values are assigned bigger values as I gets bigger, so more likely to win, increasing accuracy.



Graph 1 – Percentage for each method for 1-10 all dimensions used

KNN-I classifier in dimension subsets

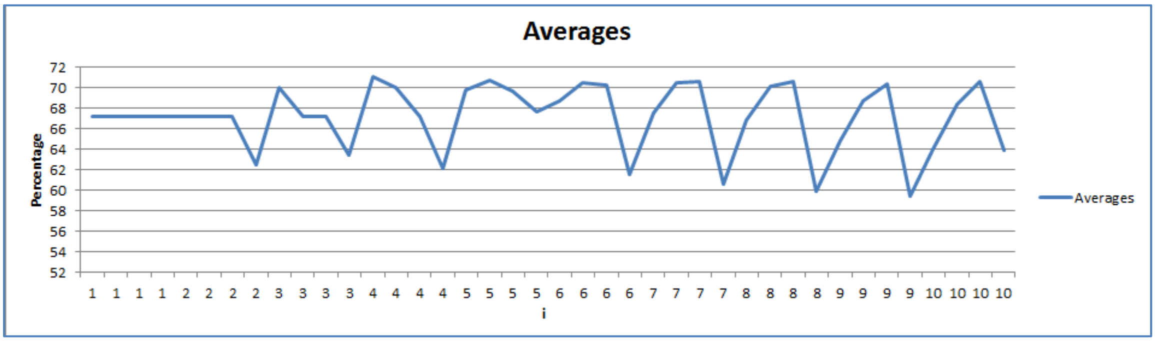
- For each I and method, all 255 dimension subsets were tested.
- Test class printOutAverages() and printOutBestPerI().



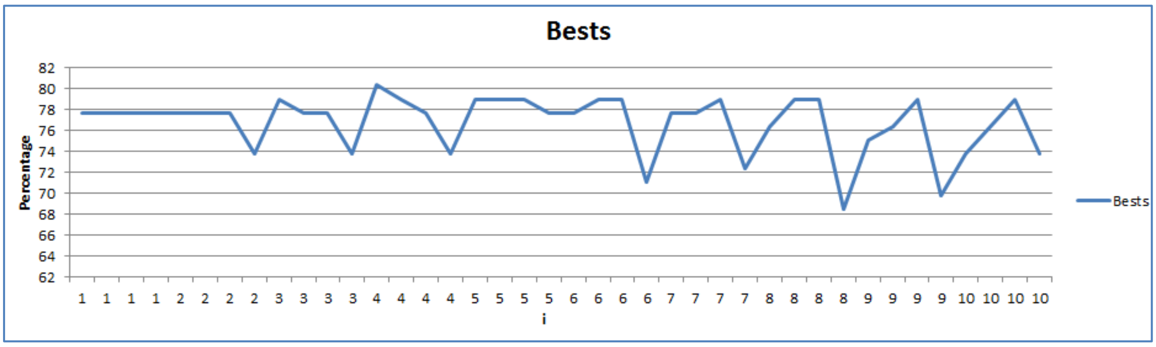
KNN	Method	Average	Best	Dimension subset used to achieve best
1	1	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	2	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	3	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	4	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
2	1	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	2	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	3	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	4	62.43	73.68	SD Depth/IQR Depth/Median Temp/IQR Temp/
3	1	69.97	78.95	Median Depth/SD Depth/IQR Temp/
	2	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	3	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	4	63.35	73.68	Mean Depth/ Median Depth/SD Depth/SD Temp/
4	1	71.00	80.26	Median Depth/SD Depth/IQR Temp/
	2	69.97	78.95	Median Depth/SD Depth/IQR Temp/
	3	67.09	77.63	Mean Depth/Median Temp/IQR Temp/
	4	62.11	73.68	Mean Depth/SD Depth/Median Temp/SD Temp/
5	1	69.67	78.95	Median Temp/IQR Temp/
	2	70.63	78.95	Median Depth/SD Depth/IQR Temp/
	3	69.63	78.95	Mean Depth/SD Depth/Mean Temp/Median Temp/SD Temp/IQR Temp/
	4	67.62	77.63	Mean Depth/SD Depth/Mean Temp/SD Temp/
6	1	68.71	77.63	Mean Depth/SD Depth/Mean Temp/Median Temp/SD Temp/IQR Temp/
	2	70.37	78.95	Median Depth/SD Depth/IQR Temp/
	3	70.22	78.95	Median Depth/SD Depth/IQR Temp/
	4	61.55	71.05	Mean Depth/SD Depth/Mean Temp/Median Temp/SD Temp/
7	1	67.52	77.63	Median Depth/Median Temp/
	2	70.38	77.63	Median Depth/SD Depth/IQR Depth/Mean Temp/Median Temp/IQR Temp/
	3	70.54	78.95	Median Depth/SD Depth/IQR Temp/
	4	60.56	72.37	Median Depth/SD Depth/Mean Temp/SD Temp/IQR Temp/
8	1	66.82	76.32	Median Depth/Median Temp/
	2	70.06	78.95	Mean Depth/Mean Temp/Median Temp/SD Temp/IQR Temp/
	3	70.55	78.95	Median Depth/SD Depth/IQR Temp/
	4	59.88	68.42	Mean Temp/IQR Temp/
9	1	64.75	75.00	Median Depth/
	2	68.64	76.32	Median Depth/Median Temp/
	3	70.32	78.95	Median Depth/SD Depth/IQR Temp/
	4	59.42	69.74	Mean Temp/IQR Temp/
10	1	63.96	73.68	Median Depth/
	2	68.34	76.32	Median Depth/Median Temp/
	3	70.50	78.95	Median Depth/SD Depth/IQR Temp/
	4	63.85	73.68	Median Depth/Mean Temp/SD Temp/IQR Temp/

Table 3 – KNN-I Dimension Subsets results

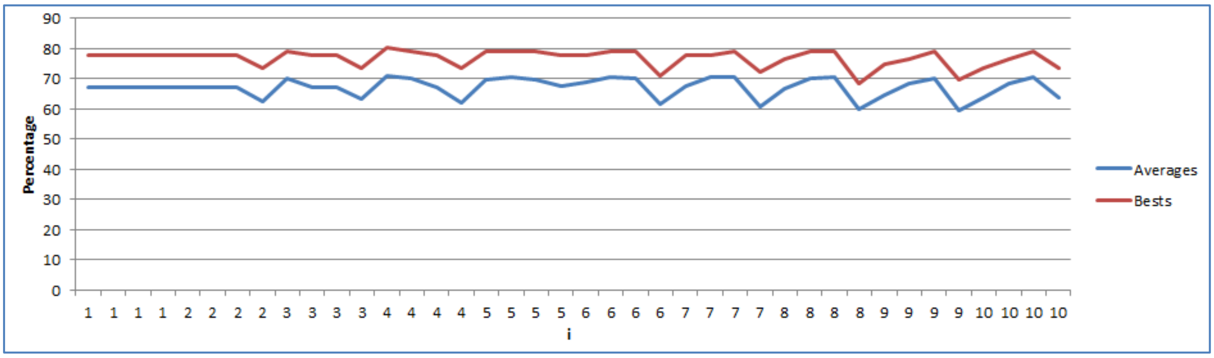
Each I shown for the 4 different methods.



Graph 2 – Average accuracy for I 1-10 and all polling methods



Graph 3 – Best accuracy for I 1-10 and all polling methods



Graph 4 – Best and Average accuracies for I 1-10 and all polling methods

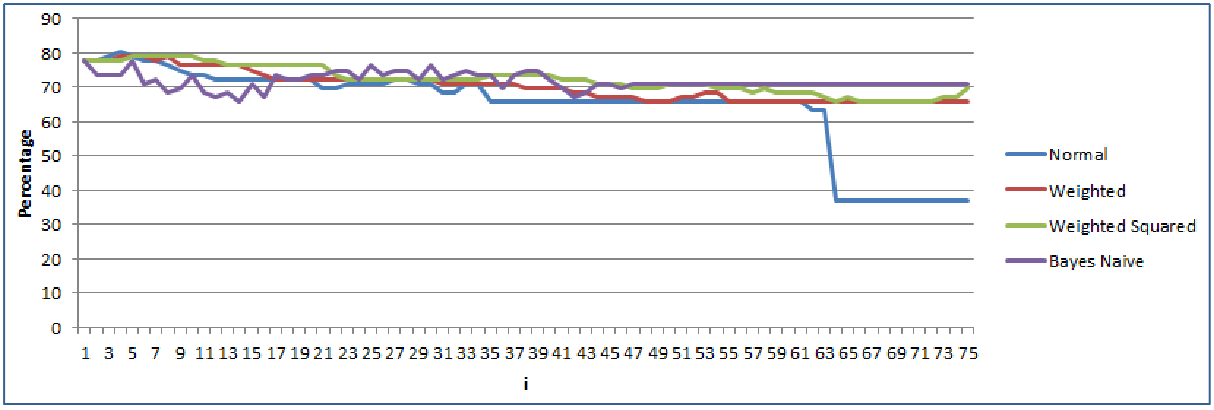
Comments:

Dimension	Appearances in best dimensions
Mean Depth	17
Median Depth	20
SD Depth	19
IQR Depth	2
Mean Temp	10
Median Temp	22
SD Temp	10
IQR Temp	30

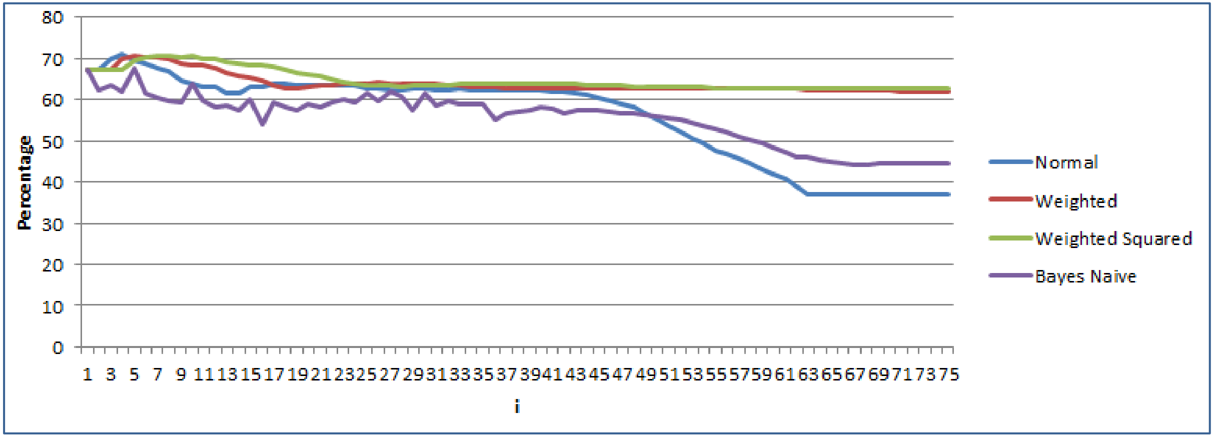
- IQR Temp most influential dimension.
- IQR Depth least influential dimension.
- 80.26% achieved using 3 of 4 most influential dimensions.
- Bayes Classifier still lower than other methods (Clear in graphs).
- No noticeable difference between 3 other methods (Clear in graphs).

Further analysis

Full list of raw data in Appendix 2/3



Graph 5 – Best accuracy for I 1-75 and all polling methods



Graph 6 – Average accuracy for I 1-75 and all polling methods

- Normal flat lines as I gets large as the classification with the largest representation in the test set (T) is returned as it always wins the vote (28 T's in the test set, 37.33%).
- The use of weighting provides a good accuracy even for large I as it is down to proximity to the sample point, not representation in the test set which hampers the normal method.

Conclusions

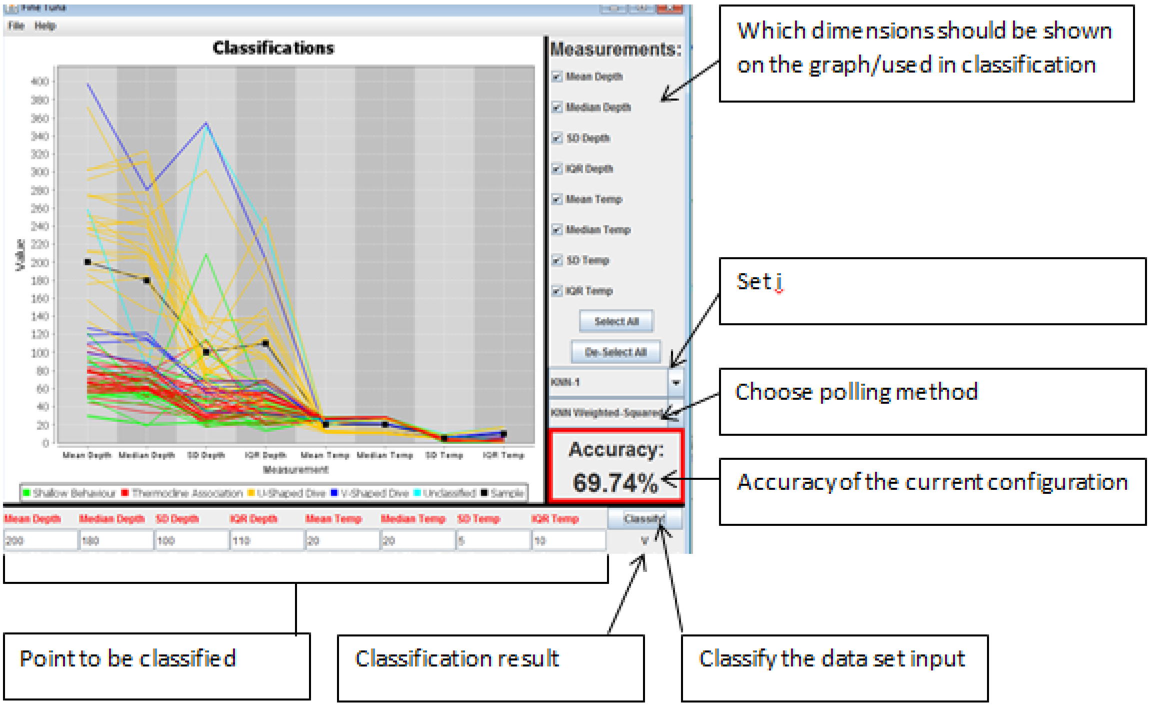
- Weighting methods are optimal as they prioritise the nearest neighbours, gives them a higher vote.
- 80% is acceptable, but still means 20% error. Other classification methods would perhaps be more appropriate and reduce the error. A higher accuracy could be achieved by altering the weighting given to certain dimensions (Such as IQR temp), however, caution should be taken as there is a possibility that characteristics of this test set (Such as the influence of IQR temp is classification) are not shared by future data. Hence future classification of new test data could result in a lower accuracy. The term Overfitting is used for this problem. Water temperature affects tuna temperature and near the surface this could be affected by the weather, so temperature measurements could be different on another test set.

Improvement Suggestions

1) Use a classifier that identifies characteristics associated with each classification. For example, there could be a threshold depth that if crossed more than 8 times in the 12 hours points towards a U/V shaped dive. Potential for a decision tree classifier.

2) Further analysis of the KNN classifier could yield interesting statistics, for example, how often each classification is predicted successfully. This could highlight alterations could be made to address any classification that is predicted correctly significantly less than other classifications, and hence improve overall accuracy.

GUI Guide



☑ Mean Depth
☑ Median Depth
☑ SD Depth
☑ IQR Depth
☑ Mean Temp
☑ Median Temp
☑ SD Temp
☑ IQR Temp

KNN-1

KNN Weighted-Squared

Accuracy: 69.74%

Mean Depth	Median Depth	SD Depth	IQR Depth	Mean Temp	Median Temp	SD Temp	IQR Temp	
200	180	100	110	20	20	5	10	V

Which dimensions should be shown on the graph/used in classification

Set i

Choose polling method

Accuracy of the current configuration

Point to be classified

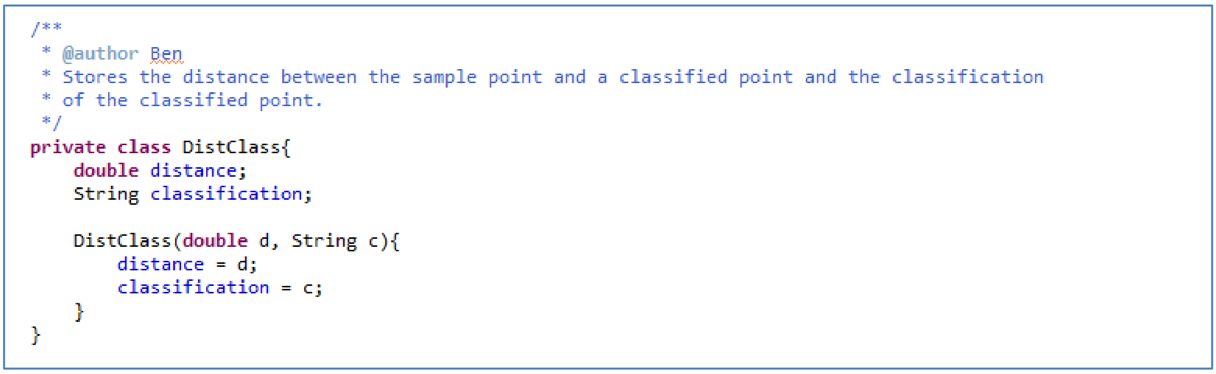
Classification result

Classify the data set input

The libraries JCommon and JFreeChart are used in Graph production. JFreeChart is free to use and available here: http://www.jfree.org/jfreechart/

Appendix

Appendix 1 - DistClass



```java
/**
 * @author Ben
 * Stores the distance between the sample point and a classified point and the classification
 * of the classified point.
 */
private class DistClass{
    double distance;
    String classification;

    DistClass(double d, String c){
        distance = d;
        classification = c;
    }
}
```

Appendix 2 – Raw Best accuracy data

I	Method 1	Method 2	Method 3	Method 4
1	77.63	77.63	77.63	77.63
2	77.63	77.63	77.63	73.68
3	78.95	77.63	77.63	73.68
4	80.26	78.95	77.63	73.68
5	78.95	78.95	78.95	77.63
6	77.63	78.95	78.95	71.05
7	77.63	77.63	78.95	72.37
8	76.32	78.95	78.95	68.42
9	75	76.32	78.95	69.74
10	73.68	76.32	78.95	73.68
11	73.68	76.32	77.63	68.42
12	72.37	76.32	77.63	67.11
13	72.37	76.32	76.32	68.42
14	72.37	76.32	76.32	65.79
15	72.37	75	76.32	71.05
16	72.37	73.68	76.32	67.11
17	72.37	72.37	76.32	73.68
18	72.37	72.37	76.32	72.37
19	72.37	72.37	76.32	72.37
20	72.37	72.37	76.32	73.68
21	69.74	72.37	76.32	73.68
22	69.74	72.37	73.68	75
23	71.05	72.37	72.37	75
24	71.05	72.37	72.37	72.37
25	71.05	72.37	72.37	76.32
26	71.05	72.37	72.37	73.68
27	72.37	72.37	72.37	75
28	72.37	72.37	72.37	75
29	71.05	72.37	72.37	72.37
30	71.05	72.37	72.37	76.32
31	68.42	71.05	72.37	72.37
32	68.42	71.05	72.37	73.68
33	71.05	71.05	72.37	75
34	71.05	71.05	72.37	73.68
35	65.79	71.05	73.68	73.68
36	65.79	71.05	73.68	69.74
37	65.79	71.05	73.68	73.68
38	65.79	69.74	73.68	75
39	65.79	69.74	73.68	75
40	65.79	69.74	73.68	72.37
41	65.79	69.74	72.37	69.74
42	65.79	68.42	72.37	67.11

43	65.79	68.42	72.37	68.42
44	65.79	67.11	71.05	71.05
45	65.79	67.11	71.05	71.05
46	65.79	67.11	71.05	69.74
47	65.79	67.11	69.74	71.05
48	65.79	65.79	69.74	71.05
49	65.79	65.79	69.74	71.05
50	65.79	65.79	71.05	71.05
51	65.79	67.11	71.05	71.05
52	65.79	67.11	71.05	71.05
53	65.79	68.42	71.05	71.05
54	65.79	68.42	69.74	71.05
55	65.79	65.79	69.74	71.05
56	65.79	65.79	69.74	71.05
57	65.79	65.79	68.42	71.05
58	65.79	65.79	69.74	71.05
59	65.79	65.79	68.42	71.05
60	65.79	65.79	68.42	71.05
61	65.79	65.79	68.42	71.05
62	63.16	65.79	68.42	71.05
63	63.16	65.79	67.11	71.05
64	36.84	65.79	65.79	71.05
65	36.84	65.79	67.11	71.05
66	36.84	65.79	65.79	71.05
67	36.84	65.79	65.79	71.05
68	36.84	65.79	65.79	71.05
69	36.84	65.79	65.79	71.05
70	36.84	65.79	65.79	71.05
71	36.84	65.79	65.79	71.05
72	36.84	65.79	65.79	71.05
73	36.84	65.79	67.11	71.05
74	36.84	65.79	67.11	71.05
75	36.84	65.79	69.74	71.05

Appendix 3 – Raw average accuracy data

I	Method 1	Method 2	Method 3	Method 4
1	67.09	67.09	67.09	67.09
2	67.09	67.09	67.09	62.43
3	69.97	67.09	67.09	63.35
4	71	69.97	67.09	62.11
5	69.67	70.63	69.63	67.62
6	68.71	70.37	70.22	61.55
7	67.52	70.38	70.54	60.56
8	66.82	70.06	70.55	59.88
9	64.75	68.64	70.32	59.42
10	63.96	68.34	70.5	63.85
11	63.28	68.21	70.07	59.53
12	62.97	67.76	69.76	58.28
13	61.48	66.51	69.18	58.46
14	61.51	65.71	68.72	57.45
15	62.99	65.33	68.57	60.03
16	63.22	64.54	68.36	54.09
17	63.76	63.54	67.87	59.35
18	63.81	62.79	67.19	58.29
19	63.42	62.75	66.53	57.31
20	63.47	63.19	66.01	58.86
21	63.46	63.34	65.6	58.1
22	63.43	63.51	65.07	59.13
23	63.29	63.78	64.33	60.09
24	63.29	63.89	63.87	59.44
25	62.79	63.98	63.61	61.44
26	62.72	64.05	63.49	59.84
27	62.42	64.03	63.32	61.78
28	62.43	63.99	63.25	60.71
29	62.54	63.93	63.35	57.49
30	62.59	63.81	63.41	61.45
31	62.48	63.69	63.5	58.37
32	62.5	63.53	63.59	59.67
33	62.53	63.35	63.68	58.99
34	62.46	63.17	63.74	59.1
35	62.39	63.03	63.83	58.75
36	62.39	62.96	63.93	55.01
37	62.44	62.86	63.92	56.79
38	62.36	62.82	63.87	56.99
39	62.23	62.83	63.84	57.61
40	62.18	62.84	63.81	58.08
41	62.02	62.79	63.81	57.96
42	61.88	62.75	63.79	56.72

43	61.57	62.7	63.69	57.54
44	61.17	62.65	63.64	57.51
45	60.59	62.65	63.54	57.51
46	59.81	62.66	63.43	57.09
47	59.09	62.65	63.34	56.83
48	57.99	62.65	63.23	56.67
49	56.66	62.67	63.14	56.34
50	55.18	62.64	63.11	55.91
51	53.55	62.62	63.12	55.62
52	51.99	62.61	63.09	55.03
53	50.69	62.63	63.01	54.35
54	49.33	62.68	62.93	53.53
55	47.76	62.66	62.84	52.74
56	46.79	62.65	62.77	52.03
57	45.8	62.65	62.74	51.16
58	44.47	62.65	62.69	50.22
59	43.17	62.63	62.7	49.49
60	42.06	62.6	62.75	48.45
61	40.9	62.59	62.73	47.17
62	38.72	62.53	62.69	46.06
63	37	62.51	62.7	45.9
64	36.83	62.47	62.68	45.44
65	36.84	62.44	62.66	45.08
66	36.84	62.38	62.63	44.43
67	36.84	62.33	62.63	44.29
68	36.84	62.29	62.61	44.29
69	36.84	62.25	62.61	44.41
70	36.84	62.18	62.62	44.44
71	36.84	62.12	62.62	44.47
72	36.84	62.05	62.62	44.47
73	36.84	62	62.62	44.47
74	36.84	61.96	62.61	44.47
75	36.84	61.91	62.63	44.47