Project 3

Purushartha Singh

March 16, 2021

Abstract

The project discusses two main methods of feature selection to reduce a problem with a high number of dimensions to the most relevant dimensions.

Contents

1 Introduction

Simply put, feature selection is the process of choosing a subset of the features from the original data according to certain criteria using Filter methods (choosing features using criteria independent of the original problem) or wrapper methods (Solving discriminant problem for each subset of features and selecting the best). It is a part of the feature extraction problem for datasets which have high number of features where classifying using all the features is highly inefficient and would lead to over fitting to the training data due to curse of dimensionality. In contrast, selecting a few of the more helpful features will greatly improve the performance and lead to better results. [1]

2 Objective

The aim of this project is feature selection for the classification problem using the high dimensional Taiji dataset provided. The two primary components of the procedure are:

1. search strategy : Identify the subset of most relevant features out of the provided features

2. evaluation function: Measure the discriminative power of the selected feature subset using filter and wrapper methods of feature selection.

3. Effect of M and N: Effect of the number of frames before and after the data which are labelled for that class on the overall classification

3 Approach

3.1 Data

The data used for the assignment was the new Taiji dataset. The dataset comprised of 1962 features and a total of 21919 data entries for each feature. It must be noted that the last element of the feature set is the frame time which was not relevant to the classification (outside of M and N value) and so, was omitted from the classification process giving a total of 1961 features.
The data points consisted of observations of 10 individuals and this was used to our advantage by using the leave one out methodology for creating 10 sets of training and testing data each containing 19918 training data points and 2001 testing data points. Further description of the features can be found in the project description.

3.2 Methodology

The following sections cover a discussion of the two primary methods used for feature selection:

3.2.1 Filter method

Definition: Given a data set, select a subset m from n features where m¡¡n such that the value of criterion function is optimized over m.
Filter methods are ”based on performance evaluation metric calculated directly from the data, without direct feedback from predictors that will finally used on the data with reduced number of features.”[1]. The two primary methods of filtering discussed in class were variance ratio and augmented variance ratio. The code for the project utilizes augmented variance ratio (AVR) to find the top 2 per cent of the entire feature set which is then passed onto the wrapper method for further selection. The evaluation uses the following formula as the metric such that the features with the highest value of the AVR metric are shortlisted. The AVR for a feature is given by:

$$AVR(F) = \frac{Var(S_F)}{\frac{1}{C}\sum_{k=1,..,C}\frac{Var_k(S_F)}{min_{m\neq k}|\mu_k(S_F)-\mu_m(S_F)|}} \tag{1}$$

such that S_F is the feature being evaluated, $Var(S_F)$ is the variance across the feature, $Var_k(S_F)$ is the variance across the feature for class k, $\mu_k(S_F)$ is the mean across the feature for class k, and $\mu_m(S_F)$ is the mean across the feature for class m such that m is not the same class as k. C is the total number of classes.

Notice that the criteria used for evaluation is independent of the specific problem. The examples are information-theory based measure such as the variance ratio, etc. and the the information-theoretic measures find the correlation, mutual information or information gain of the different feature subsets similar to how Fisher method does.
Once the AVR has been found for each feature, the features are ranked and the highest scoring features are selected.

Advantages

1. **Fast execution:** Only a subset of the features are used for the rest of the problem, thus, speeding up the computation.
2. **Generality:** Since the evaluation criteria is not data-specific, this method can be generalized over data sets.

Disadvantages

1. It does not reveal any information about the selected feature aside from ability to split the dataset along the classes for the training dataset. As such, it may not be useful for making informed decisions and the non specific way can lead to loss of information.

3.2.2 Wrapper method

Definition: In this method, the features are selected on the basis of quantitative rates. More specifically, the wrapper method implies solving of discriminant problem for each subset of features and selecting the best classification rates. For each subset of features selected by the wrapper methods should separate the class distribution as much as the

original class distribution.
The search strategy is the systematic procedure to choose there candidate feature subsets. These include Complete, greedy and randomized algorithms. The approach taken in the project is the complete approach using Sequential Forward Selection. Since the input is limited to the top AVR selections from the filter method, the NP complete nature of the problem is circumvented by keeping the input size small. The following algorithm is used to recursively find the best fitting accuracy on the training method (using Linear discriminant classification method):

1. Start with an empty set P and a set of features Q
2. Select the next best feature $x^+ = argmax[Q(Y_k) + x]$ where Yth feature of Q is the evaluation criterion.
3. Update $P_{k+1} = P_k + x^+$, $Q_{k+1} = Q_k - x^+$and $k = k + 1$
4. Jump to step 2 until the maximum value of x^+ has been achieved and no further addition of features increases the classification accuracy on the training set $\forall x^+ \in Q_{k+1}$.
5. Return P_{k+1}

Advantages

1. **Accuracy:** The method finds the best set of features from the given set to maximize the training classification accuracy.
2. **Avoids over fitting:** Uses cross-validation for predictive accuracy

Disadvantages

1. **Slow execution:** Given the NP complete run time, the problem can be very resource intensive if the input is not managed to very small set of features.
2. **Lack of generality:** The approach is highly dependent on the classifier and feature subset. In fact, during the project, using different classifiers lead to significant increase and decrease in the accuracy of the forwarded features.

3.3 Workflow

As requested, the workflow of the project has been provided below in fig 1:

4 Results

4.1 Feature Selection

This analysis was done with default values of M = 100 and N = 20. The Filter method selected top 1% of the features based on AVR. From the resulting classification, an average accuracy of 0.6958 was achieved over 10 iterations using all but one method to make

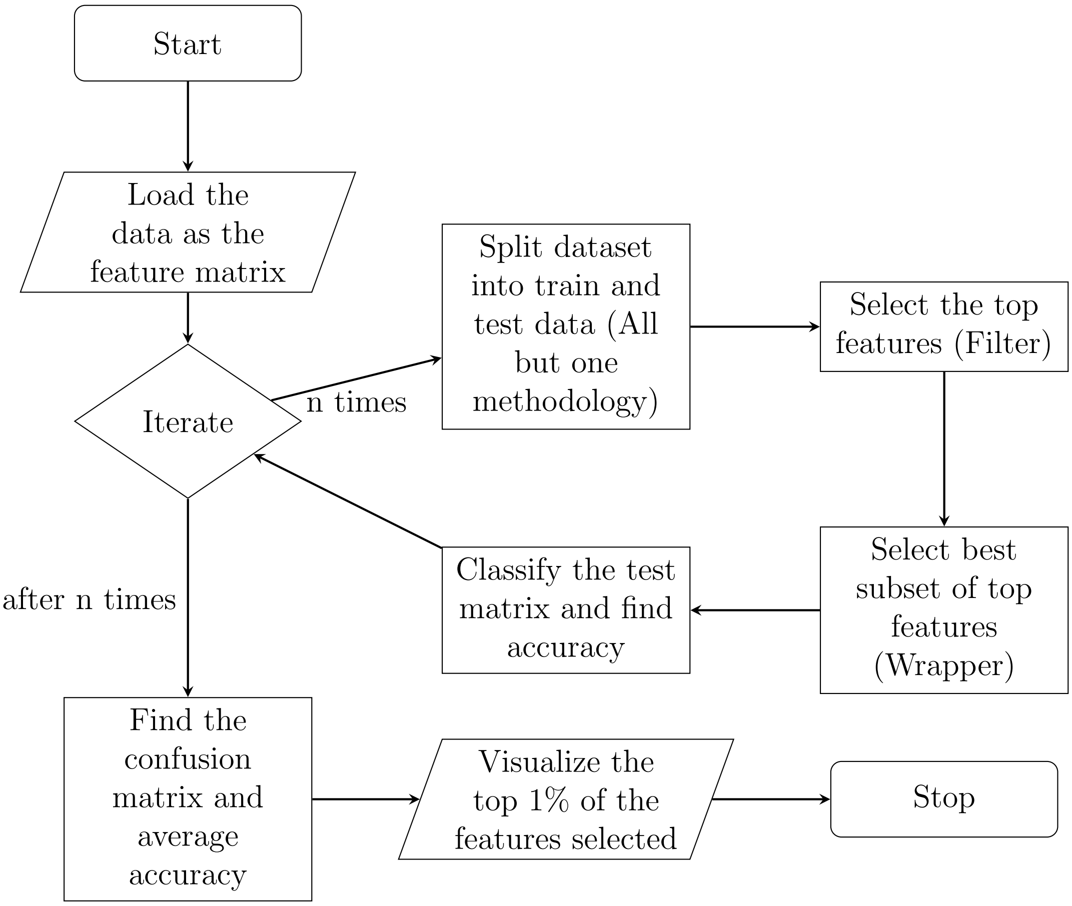

Start
Load the data as the feature matrix
Iterate
n times
Split dataset into train and test data (All but one methodology)
Select the top features (Filter)
Select best subset of top features (Wrapper)
Classify the test matrix and find accuracy
after n times
Find the confusion matrix and average accuracy
Visualize the top 1% of the features selected
Stop

Figure 1: The basic workflow of the code. Details of each step can be found in the methodology section

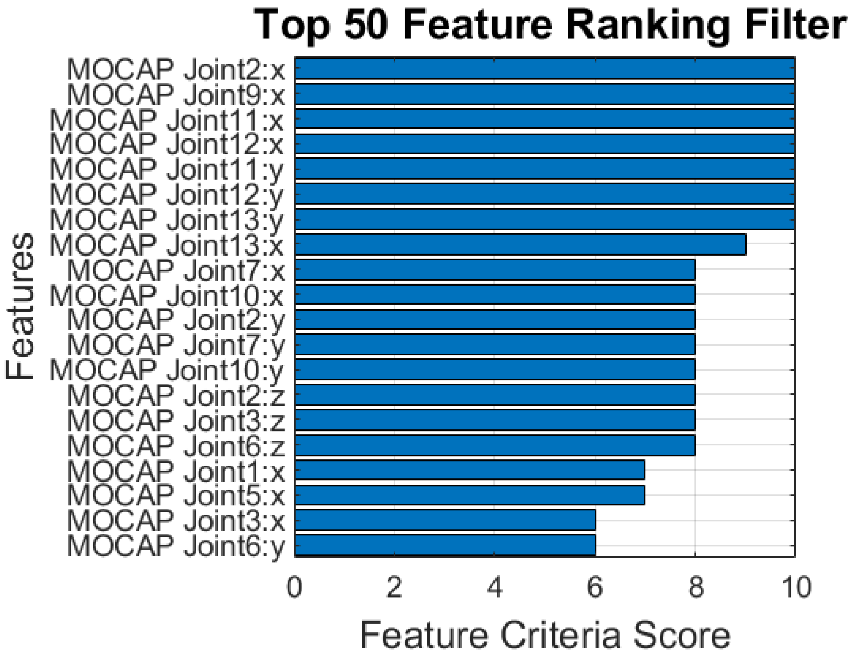


Figure 2: Frequency of appearance of features over 10 iterations in the filter method of feature selection

the training and testing datasets. The 20 most used features in this task for the filter and wrapper methods are shown in figure 2 and 3 respectively. Note that for all the classifications, Linear discriminant classification was used.

4.2 Classification Analysis

4.2.1 Base run

For the given parameters of M=100, N=20, K=0.01 (Filter % value of 1%), and Linear discriminant classifier, the average accuracy achieved for the run was 0.6983 with a standard deviation of 0.2740. The classification matrix and confusion matrix have been shown in fig 4 and 5 respectively. It must be noted that due to lack of space, the labels in the classification matrix have been omitted. For this run, the classification rate for each class for each iteration and the average classification rate for each class over all the iterations are shown in 6.

4.2.2 Final Run

With the values of more optimal parameters known, a final run with M=60, N=400, K=0.05, and Ensemble classifier got average classification accuracy of **0.8342**. with average standard deviation of 0.1538. The classification matrix and confusion matrix have been shown in fig 7 and 8 respectively.For this run, the classification rate for each class for each iteration and the average classification rate for each class over all the iterations are shown in 9. Further optimization may be done to improve the classification accuracy further.

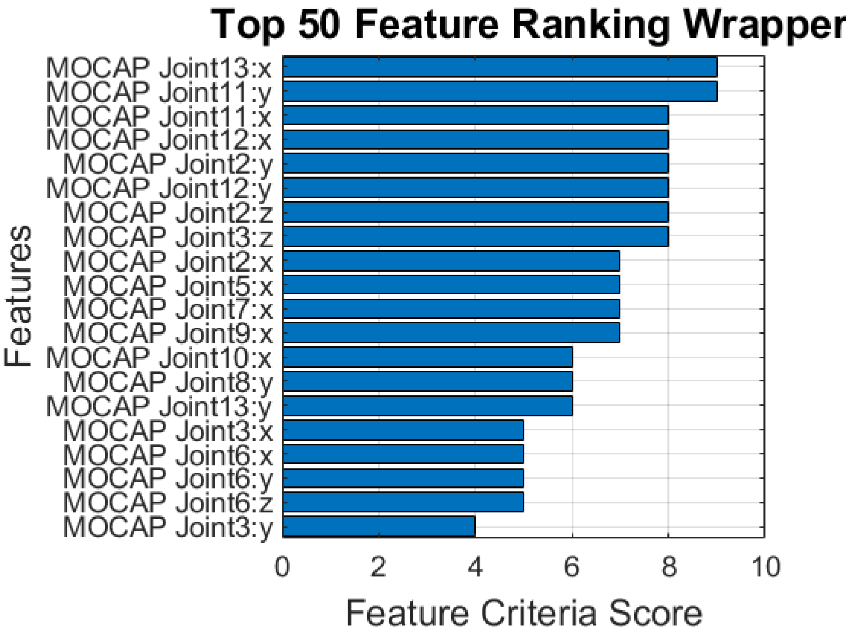
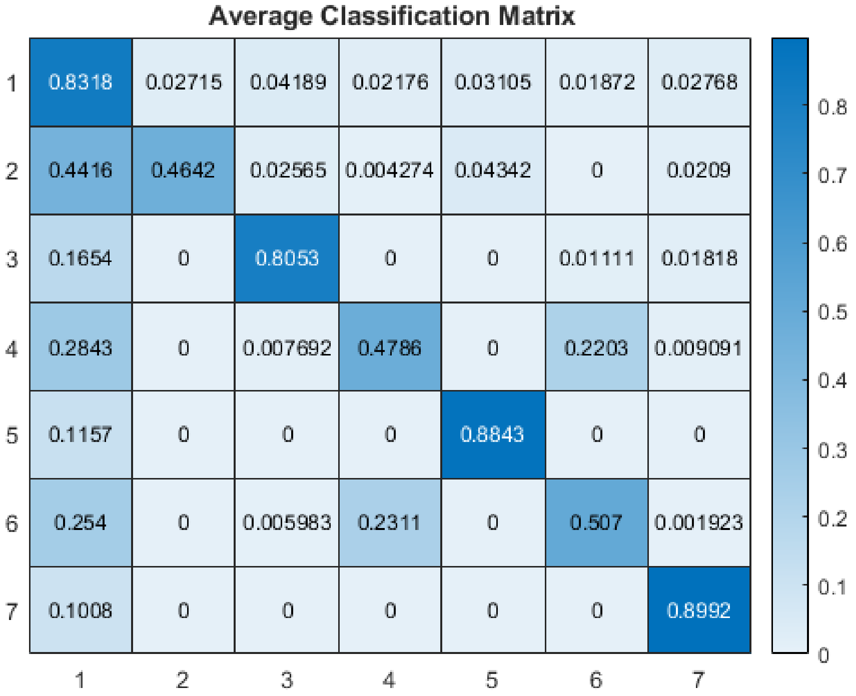
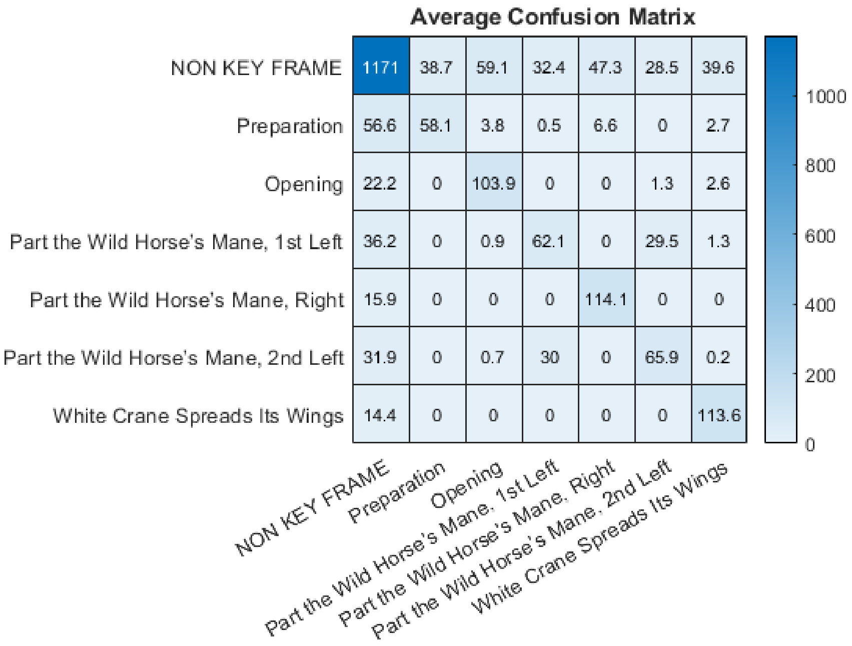


Figure 3: Frequency of appearance of features over 10 iterations in the wrapper method of feature selection



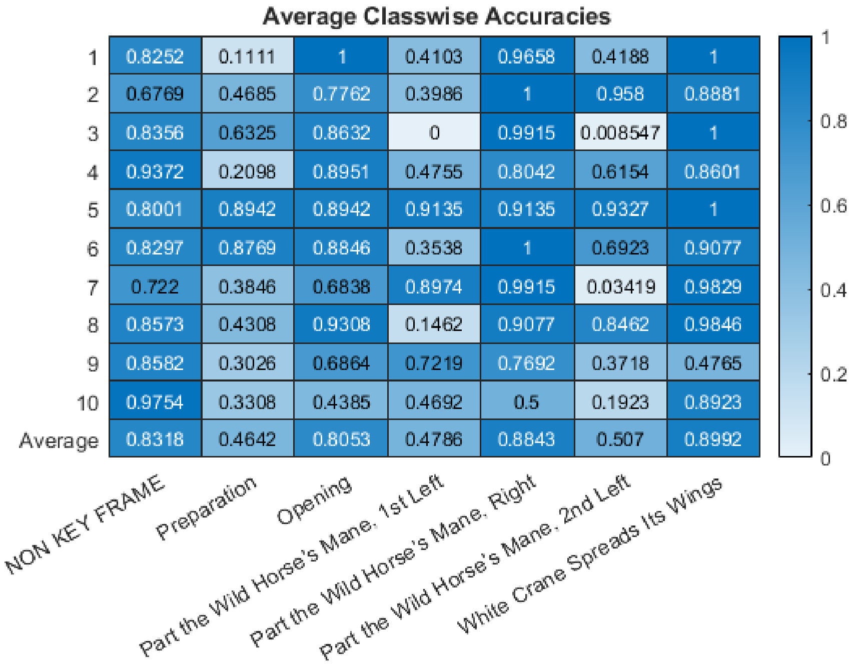
	1	2	3	4	5	6	7
1	0.8318	0.02715	0.04189	0.02176	0.03105	0.01872	0.02768
2	0.4416	0.4642	0.02565	0.004274	0.04342	0	0.0209
3	0.1654	0	0.8053	0	0	0.01111	0.01818
4	0.2843	0	0.007692	0.4786	0	0.2203	0.009091
5	0.1157	0	0	0	0.8843	0	0
6	0.254	0	0.005983	0.2311	0	0.507	0.001923
7	0.1008	0	0	0	0	0	0.8992

Figure 4: M=100, N=20, K=0.01, Classification Matrix



	NON KEY FRAME	Preparation	Opening	Part the Wild Horse's Mane, 1st Left	Part the Wild Horse's Mane, Right	Part the Wild Horse's Mane, 2nd Left	White Crane Spreads Its Wings
NON KEY FRAME	1171	38.7	59.1	32.4	47.3	28.5	39.6
Preparation	56.6	58.1	3.8	0.5	6.6	0	2.7
Opening	22.2	0	103.9	0	0	1.3	2.6
Part the Wild Horse's Mane, 1st Left	36.2	0	0.9	62.1	0	29.5	1.3
Part the Wild Horse's Mane, Right	15.9	0	0	0	114.1	0	0
Part the Wild Horse's Mane, 2nd Left	31.9	0	0.7	30	0	65.9	0.2
White Crane Spreads Its Wings	14.4	0	0	0	0	0	113.6

Figure 5: M=100, N=20, K=0.01, Confusion Matrix



	NON KEY FRAME	Preparation	Opening	Part the Wild Horse's Mane, 1st Left	Part the Wild Horse's Mane, Right	Part the Wild Horse's Mane, 2nd Left	White Crane Spreads Its Wings
1	0.8252	0.1111	1	0.4103	0.9658	0.4188	1
2	0.6769	0.4685	0.7762	0.3986	1	0.958	0.8881
3	0.8356	0.6325	0.8632	0	0.9915	0.008547	1
4	0.9372	0.2098	0.8951	0.4755	0.8042	0.6154	0.8601
5	0.8001	0.8942	0.8942	0.9135	0.9135	0.9327	1
6	0.8297	0.8769	0.8846	0.3538	1	0.6923	0.9077
7	0.722	0.3846	0.6838	0.8974	0.9915	0.03419	0.9829
8	0.8573	0.4308	0.9308	0.1462	0.9077	0.8462	0.9846
9	0.8582	0.3026	0.6864	0.7219	0.7692	0.3718	0.4765
10	0.9754	0.3308	0.4385	0.4692	0.5	0.1923	0.8923
Average	0.8318	0.4642	0.8053	0.4786	0.8843	0.507	0.8992

Figure 6: M=100, N=20, K=0.01, Class wise accuracy and average for all iterations

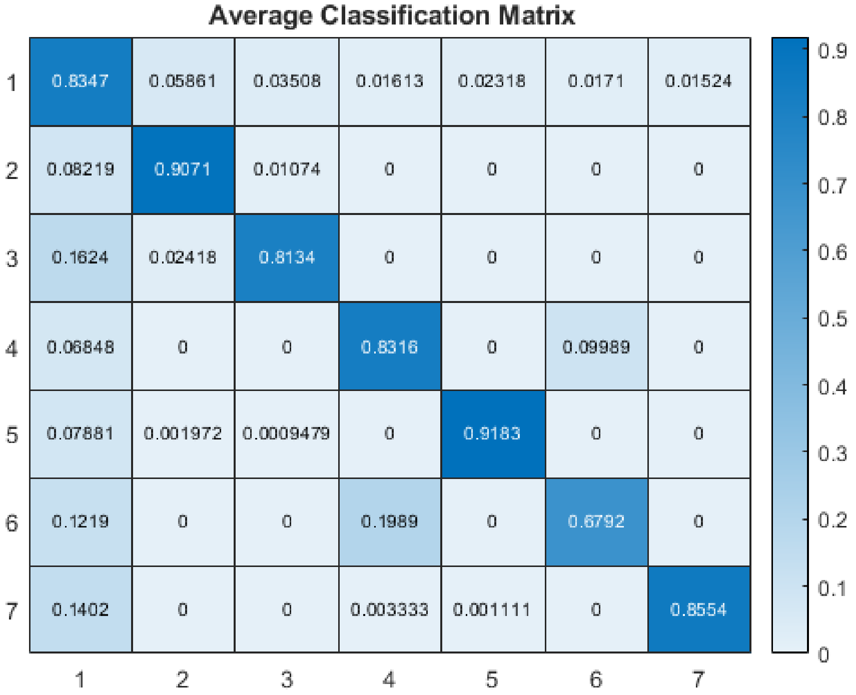


Figure 7: M=60, N=400, K=0.05, Ensemble Classification Matrix

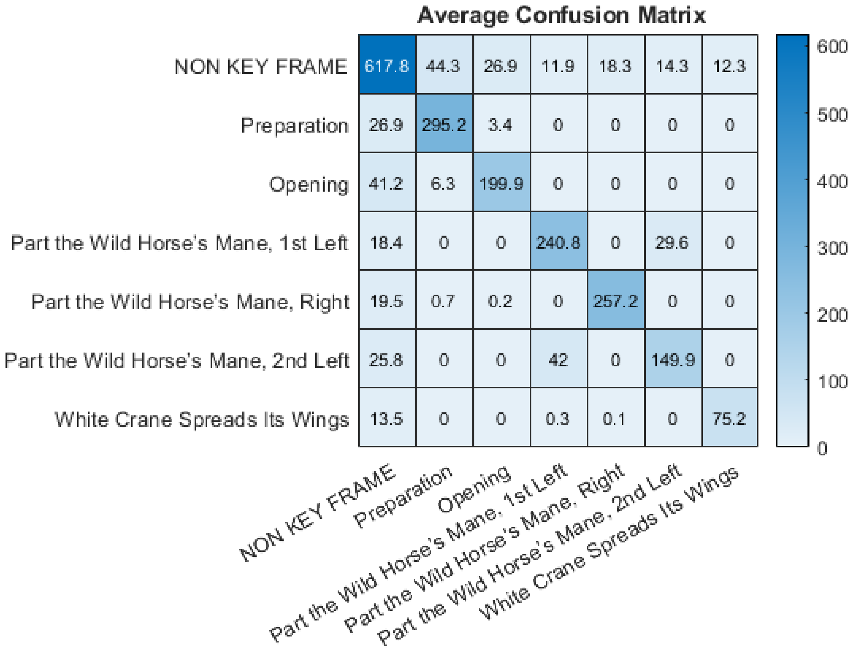
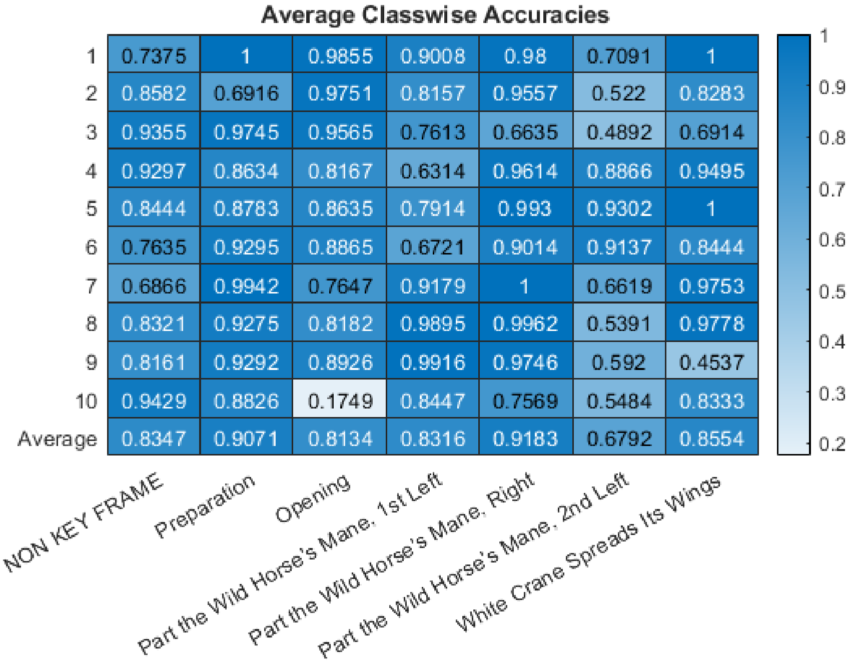


Figure 8: M=60, N=400, K=0.05, Ensemble Confusion Matrix



Average Classwise Accuracies

	NON KEY FRAME	Preparation	Opening	Part the Wild Horse's Mane, 1st Left	Part the Wild Horse's Mane, Right	Part the Wild Horse's Mane, 2nd Left	White Crane Spreads Its Wings
1	0.7375	1	0.9855	0.9008	0.98	0.7091	1
2	0.8582	0.6916	0.9751	0.8157	0.9557	0.522	0.8283
3	0.9355	0.9745	0.9565	0.7613	0.6635	0.4892	0.6914
4	0.9297	0.8634	0.8167	0.6314	0.9614	0.8866	0.9495
5	0.8444	0.8783	0.8635	0.7914	0.993	0.9302	1
6	0.7635	0.9295	0.8865	0.6721	0.9014	0.9137	0.8444
7	0.6866	0.9942	0.7647	0.9179	1	0.6619	0.9753
8	0.8321	0.9275	0.8182	0.9895	0.9962	0.5391	0.9778
9	0.8161	0.9292	0.8926	0.9916	0.9746	0.592	0.4537
10	0.9429	0.8826	0.1749	0.8447	0.7569	0.5484	0.8333
Average	0.8347	0.9071	0.8134	0.8316	0.9183	0.6792	0.8554

Figure 9: M=60, N=400, K=0.05, Ensemble Class wise accuracy and average for all iterations

4.3 EC: Use of different classifiers and ensemble learning

Use of different classifiers lead to pretty much the same results with the base parameters of M=100, N=20, K=0.01. Here is a list of the accuracy results for KNN, and decision tree classifier:

- K- Nearest Neighbors: 0.7068
- Decision Tree: 0.6412
- Naive Bayes: 0.7026
- Linear Discriminant (Base): 0.6983

Ensemble learning is used with a few parameters to improve the performance compared to these three methods to **0.7295**. This can be attributed to the following parameters:

- First, the bag method is used which is a bootstrap method. This allows for artificial increase of sample data without need for more actual observations.
- The expected value of prior is changed to the empirical values in the training data. This allows for a weight to be assigned to each expected value based on the probability of an item appearing in the provided data, thereby reducing the weight of data of classes which appear more often and boosting the data which comes from classes with fewer datapoints.
- Number of Bins is increased to 50 to improve the speed of the classifier since it allows for multiple values to be calculated in parallel making the process more time efficient.

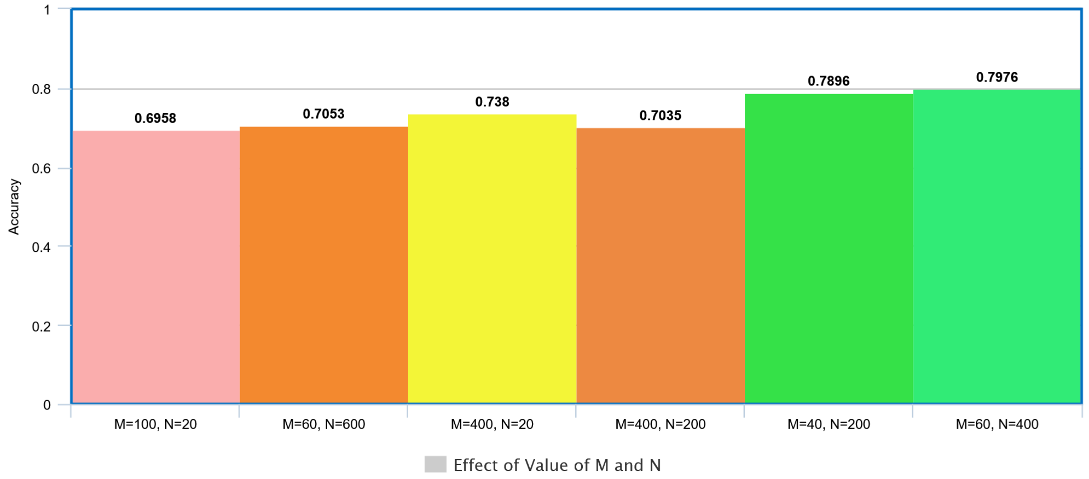


Figure 10: Effect of M and N values on the classification accuracy. K=0.01

4.4 Effect of M and N

M stands for the number of frames before the pose and N stands for number of frames after the pose. These give the data appearing at frame values in M-N range the assigned class for the pose during the assignment of classes to the data. The values of M and N seem to have a significant impact on the average accuracy of the classification as can be seen in fig 10. For uniformity, the value of K is kept at 0.01 and Linear discriminant classifier is used. As can be seen in the figure, value of accuracy increases by 10% by tuning the values of M and N.
The reason for this improvement can be due to the fact that increasing the number of labelled frames would decrease the large number of unclassified frames which often end of causing the confusion matrix to skew towards it. Having a more balanced dataset gives better classification accuracy.

5 Conclusion

Feature selection is important to reduce the computational efforts and space for working with large data sets. It helps identify features that are most useful and thus, improve the overall accuracy of the algorithm.

Given the novelty of the dataset, it was a very interesting experience having to learn on the go about different parameters and how multiple aspects of reducing dimensionality must be looked at. In addition, this was my first introduction to data specific parameters such as M and N and I was very surprised to look at the drastic effect of the two on the overall accuracy of the data.

This was also a very interesting look at how dimensionality reduction ends up causing a lot of data loss at times but at the same time, keeping the extra dimensions often results in poorer results. Overall, I was a bit sad that the time period for this project was short and as a result, I was not able to be more thorough with testing and parameter tuning. It would have also been very interesting to have attempted some of the other extra credit activities such as data visualization and running this on the bigger dataset which i was unable to do due to lack of time.

References

[1] Christopher M. Bishop *Pattern Recognition and Machine Learning.* 2006 2, 3