Natural Disaster and Weather Prediction System

Milind Kamath, Priyanka Punjabi, Vineet Kamat
Department of ComputerScience
Rochester Institute of Technology
Rochester, NY 14623
{mk6715;pp4762;vvk1199}@rit.edu

ABSTRACT

The ENSO cycle is a key contributor to the US weather cycle. El Nino/Southern Oscillation (ENSO) cycles is based on a recurring climate pattern over the central and eastern Pacific Ocean. This climate variation occurs across the region in between Indonesia and USA. The purpose of this project is to create a warning system by finding a pattern which leads to or might lead to natural calamities.

1. INTRODUCTION

In the year 1982-83, the world witnessed the most extreme ENSO cycles. This extremity led to heavy rains and snow in California while the northern and north eastern regions received below average snowfall. Also, the areas like Gulf of Mexico witnessed fewer hurricanes and tropical storms. The ENSO cycles were one of the contributors.[1] This cycle lasts between 3 to 7 years and is divided into 3 phases namely:[6]
El Nino: The temperature above the eastern and central pacific region is more than the average sub surface temperature (SST). This occurs when low level surface winds that move from east to west (eastern winds) weaken as rains move from west to east. This affects the lower altitude winds which tend to change direction and convert to western winds. Thus, warmer the ocean temperature, stronger the EL Nino phase.
La Nina: The temperature above the eastern and central pacific region is less than the average sub surface temperature (SST). As rainfall increases over the Indonesian region, wind moves east to west which indirectly strengthens the eastern winds. Thus, stronger the eastern winds, stronger is the La Nino phase.
Neutral: The temperature above the eastern and central pacific region is near to the average sub surface temperature (SST).

The above 3 phases have a considerable effect on the polar jet stream from the Arctic, Pacific jet stream, subtropical jet stream and the Atlantic jet stream as can be seen in the below image

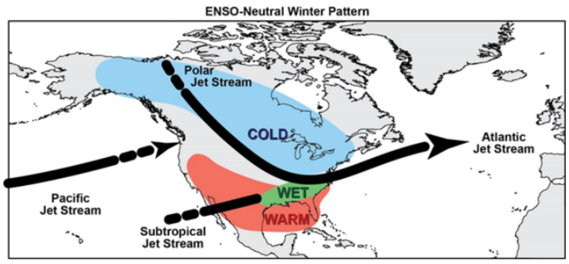


Figure 1: Effect of the phases [6]

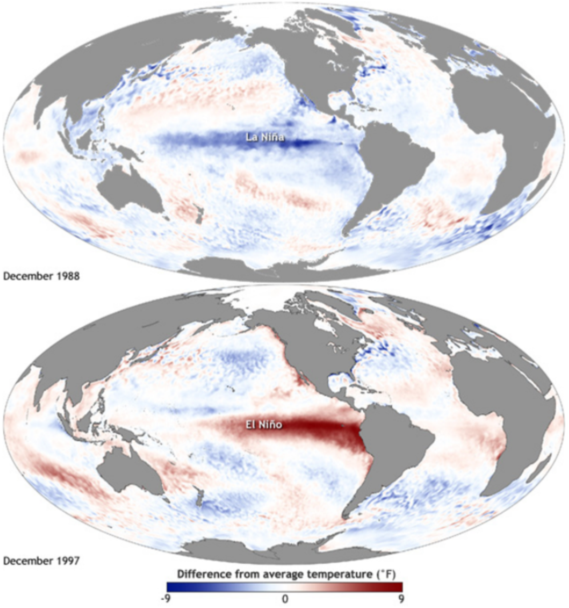


Figure 2: Phases [6]

With the help of this data it is possible to predict the climate for the next 1 or 2 years. But in the year 1982-83 this prediction was proved false since the models were unable to detect extreme anomalies in climate. This led to the development of the Tropical Atmosphere Ocean (TAO) array by Tropical Ocean Global Atmosphere (TOGA) program. TAO consists of 70 buoys capable of measuring air temperature, relative humidity, surface winds, sea surface temperatures, subsurface temperatures (upto 500 meters depth), currents, rainfall and solar radiation on a real time basis.[4]

With the given dataset, the project aims to find a suitable data mining algorithm that can best predict the weather. We have analysed the dataset using classification rule based PRISM, classification tree based RandomForest, clustering using K-means[5] and DBScan[2].

The overall paper has been structured as Section 2 describes the dataset and Section 3 discusses the pre processing conducted to clean the dataset. Section 4 talks about data visualization for better perspective. Section 5 summarizes the steps involved in selecting the data mining algorithm to be used, Section 6 and 7 talks about the design and implementation of clustering respectively. Section 8 states the conclusion and Section 9 gives a brief on future work that can be undertaken.

2. DESCRIPTION OF DATA-SET

The project involves managing 2 data-sets [3]:
ENSO: The enso dataset contains the following attributes:

- buoy - Buoy number
- day - Day number when the reading was noted
- latitude - Latitude of buoy in pacific ocean
- longitude - Longitude of buoy in pacific ocean
- zon.winds - Speed of Zonal Winds (Winds along the latitude)
- mer.winds - Speed of Meridional Winds (Winds along the longitude)
- humidity - Humidity of air above the surface
- air temp. - Air temperature
- s.s.temp. - Surface temperature. The surface temperature changes at a much slower rate than the air temperature.

All attributes in the ENSO dataset are numeric in nature.

TAO: The tao dataset contains the following attributes:

- obs - observation number
- year - year of observation
- month - month of observation
- day - day of observation
- date - Combination of day, month and year
- latitude - Latitude of buoy from TAO array in pacific ocean
- longitude - Longitude of buoy from TAO array in pacific ocean
- zon.winds - Speed of Zonal Winds (Winds along the latitude)
- mer.winds - Speed of Meridional Winds (Winds along the longitude)
- humidity - Humidity of air above the surface
- air temp. - Air temperature
- s.s.temp. - Surface temperature. The surface temperature changes at a much slower rate than the air temperature.

All attributes in the TAO dataset are numeric in nature.

3. DATA PRE - PROCESSING

Data Pre – processing is a technique that involves transforming a raw data into an understandable format. In this phase we create a csv file from the given .dat files, eliminating unnecessary columns, clean missing values and discretize our data.

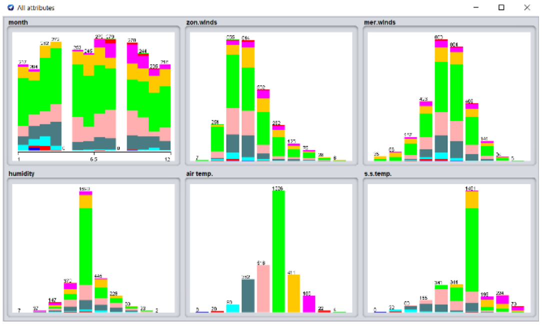
- **Creating CSV Files:** We created a csv file for the data collected using a python script. Pandas library was used to read the .dat file. The .dat file contained attributes with multiple spaces, hence, '\s+' regex expression was used as a delimiter for reading the file.
- **Eliminating Unnecessary Columns:** To decide on a predictive model which will be distributed across the buoys, we conducted a preliminary analysis on a randomly selected buoy. This randomly selected buoy had a total of 2946 instances over a period of 1986 to 1998 ie. 12 years. As we all know climate changes are seasonal, modeling climate change on a daily basis will not generate an accurate model. The training process will also be highly compute and memory consuming. Also in our case, we do not have any external data source related to yearly changes. Thus factoring the year in our model will lead to divergence and unknown information gain. It will also cause reduction in accuracy as well as model clarity. Thus we have removed irrelevant attributes number, day, year, date, etc from our modelling.
- **Missing Values:** Our dataset consisted of missing values which needed to be fixed. To do so, we calculated the mean values of similar data and replaced the missing values. In order to find similar data, we first grouped the records based on buoy number (Elnino dataset) and latitude longitude (TAO dataset) attributes. The latitude and longitude attributes had a slight deviation in their values, so in order to group the records, it was imperative to normalize these values around its mean.
  Thus our first step was to find the normalized values for latitude and longitude attributes. We executed EM clustering algorithm on attributes latitude and longitude independently and found the below results. As we can see longitude has 9 values which are represented by 9 clusters and latitude has 13 values represented by 13 clusters.

Longitude									
Attribute	0	1	2	3	4	5	6	7	8
mean	165.0104	-140.234	-110.02	-125.028	-179.557	-155.234	-95.0474	154.5264	-169.974
std. dev	0.07	1.1736	0.2194	1.5252	1.091	0.7444	0.2251	8.7295	0.1428

Latitude													
Attribute	0	1	2	3	4	5	6	7	8	9	10	11	12
mean	-4.9939	4.9997	5.0178	-0.0419	5.925	-8.0477	8.0141	5.3306	-2.0147	6.946	8.098	8.9878	1.9653
std. dev	0.0427	0.016	0.0572	0.1575	0.015	0.1048	0.0319	0.0554	0.0781	0.1282	0.1004	0.0172	0.2083

After normalizing the latitude and longitude attributes, we replace the missing values of the remaining attributes in respective lat. and long. ranges with the mean of that range. If all the values for an attribute within that range are missing, then the missing values are replaced with the global mean of that attribute. Similarly for elnino, we already have the buoy number, so the grouping was done based on that attribute.

- **Discretize Data:** Every predefined model considered in our cases works best with nominal values. Our dataset is a numeric dataset. The values are such that they vary upto 3 decimal places. If we directly convert numeric values to nominal values, then the models generated will be highly specific, sparse and eventually highly inaccurate in the evaluation phases. Thus we first discretized our model using the discretize function provided by Weka. For our preliminary analysis, we distributed our attribute values across 10 equally distant buckets. These discretized values were then converted to nominal values. Below is one of the visualizations, showing how air temp attribute is relatively significant.

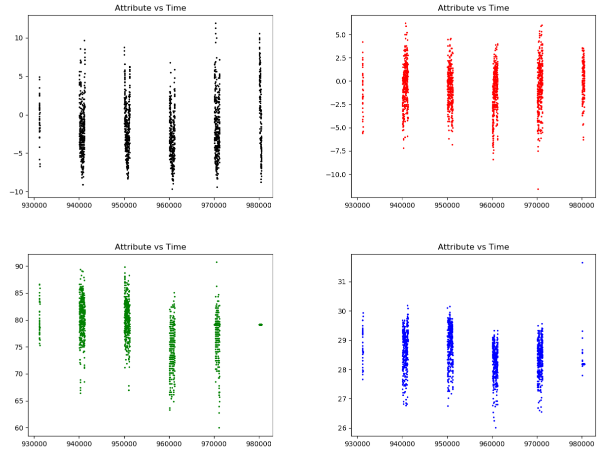


All the data pre – processing of both datasets was performed in WEKA which is very powerful language which detects missing values and also helps us to eliminate and add attributes to the dataset easily.

4. DATA VISUALIZATION

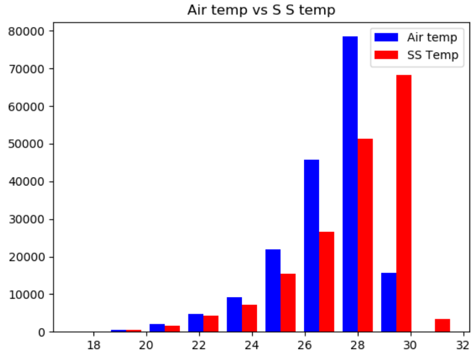
Our next step is visualizing the dataset in – order to gain meaningful insights out of it. Since the dataset is huge, it is crucial to visualize the data before actually using it to implement our data mining algorithms. We considered attributes like zonal winds, meridian winds, humidity, air temp and ss temp for evaluation.

Visualization 1: Scatter Plot of buoy 1 with zon. winds, mer. winds and humidity

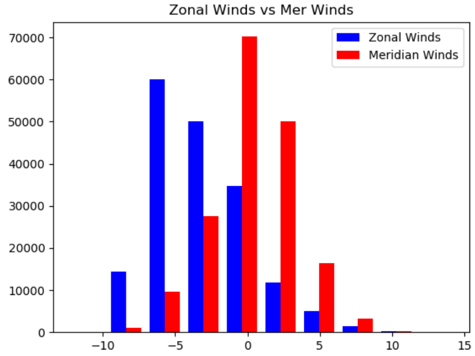


The scatter folder contains per buoy per attribute plots. Since there were 86 unique buoys and 5 attributes for consideration, 430 scatter plots were created, in order to visualize the trend per buoy per attribute.

Visualization 2: Histogram highlighting relation between air temp and ss temp.



Visualization 3: Histogram highlighting relation between zonal winds and meridian winds.



5. DATA MINING

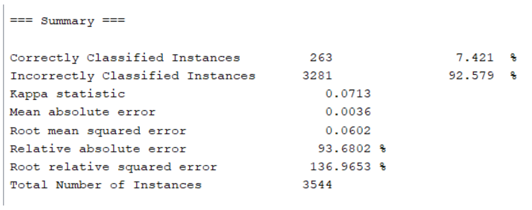
The TAO dataset is from readings from buoys over many years, it records various readings like meridian and zonal winds, humidity and air temperature etc. The dataset doesn't

have a label as to what these readings actually represent, and hence the idea of clustering values that have similar attributes together. Using EL-nino dataset, which is over a period of 15 days which represent heavy winds and mountain snow and such unusual behavior, the clustering can be studied.
The TAO dataset has been analysed using clustering and classification methods. The results of this analysis is based on parameters like incorrectly classified instances, Silhoutte Coefficient and Accuracy. We have used K means and DBScan for cluster based analysis, rule based classification using PRISM and tree based classification using RandomForest.
Our initial setup was to use the complete dataset. Since the dataset was numeric, it was required to convert them to nominal values. Our first step was to discretize the value and then convert to numeric values. As shown in the last report, due to sparsity the accuracy of the model was not up to the mark. The next approach was to group the dataset based on seasons. The reason for this step was to bring instances close to each other. But with this approach we faced the curse of dimensionality problem. The overall performance was still below mark as it can be seen in subsections.
From the pre-analysis shown in the previous section, we can see air temp and ss temp have the highest correlation. Thus we considered only these two attributes for clustering and classification purpose and a drastic change in the accuracy and silhoutte coefficient was observed.

5.1 Classification

We have discretized all attribute values in 4 bins. The below results are for a single buoy selected randomly. We analysed the same using Weka and we found the below results.

1. **Prism**
   Target Class - SS Temp As you can clearly see more than 90% of the instances were incorrectly classified. Hence use of Classification Rule based using PRISM with ss temp cannot be used for predicting the weather.
   Classification Rule based using PRISM with SS Temp target class

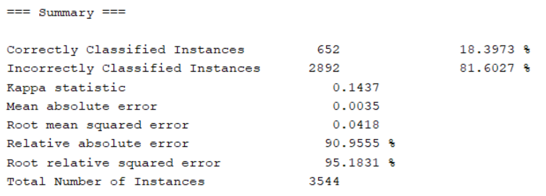


```
=== Summary ===

Correctly Classified Instances         263                7.421  %
Incorrectly Classified Instances      3281               92.579  %
Kappa statistic                          0.0713
Mean absolute error                      0.0036
Root mean squared error                  0.0602
Relative absolute error                 93.6802 %
Root relative squared error            136.9653 %
Total Number of Instances             3544
```

   Target Class - Month
   As you can clearly in Fig 3. see more than 70% of the instances were incorrectly classified. Hence use of Classification Rule based using PRISM with month cannot be used for predicting the weather.

   To improve accuracy we removed all attributes other than air temp and ss temp with month as the target class and found the results as shown in Fig 4 in Weka. But as you can see below there was no improvement with this method. Also all instances got classified as belonging to the first season.

2. **RandomForest**
   Target Class - SS Temp
   As you can clearly see more than 80% of the instances were incorrectly classified. Eventhough this is better than PRISM, we cannot use RandomForest for predicting the weather.
   RandomForest with SS Temp target class



```
=== Summary ===

Correctly Classified Instances         652               18.3973 %
Incorrectly Classified Instances      2892               81.6027 %
Kappa statistic                          0.1437
Mean absolute error                      0.0035
Root mean squared error                  0.0418
Relative absolute error                 90.9555 %
Root relative squared error             95.1831 %
Total Number of Instances             3544
```

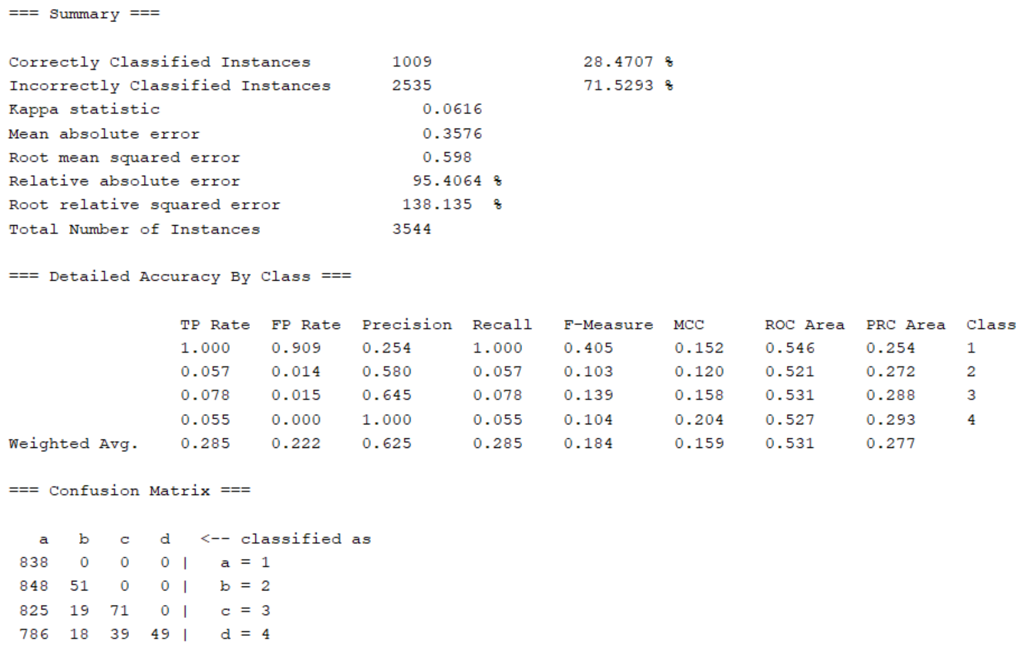
   Target Class - Month
   As you can clearly in Fig 5 almost 50% of the instances were incorrectly classified. Even though data mining using Classification Tree based RandomForest with month produced better results, other method should be taken into consideration before finalizing the model to be used for predicting the weather.
   To improve accuracy we removed all attributes other than air temp and ss temp with month as the target class and found the results as shown in Fig. 6 in Weka. As it can be clearly see the number of incorrectly classified instances increased.

5.2 Clustering

The TAO dataset is used first. The data is cleaned and missing data is fixed and as mentioned before, this dataset contains readings of many buoys over a period of many years. Representation of what exactly those features represent is not present in the dataset. Such unlabelled data needs unsupervised learning techniques. Hence clustering was used here as one of the methods to analyze the data. El-nino data was then used to label the clustered data based on the values these points went into.
Clustering and classification sometimes go hand in hand, and that is the idea used here. Clustering the TAO dataset and using EL-Nino to label the clusters is the motivation here. A more promising result was found using clustering and its implementation has been explained in further sections.

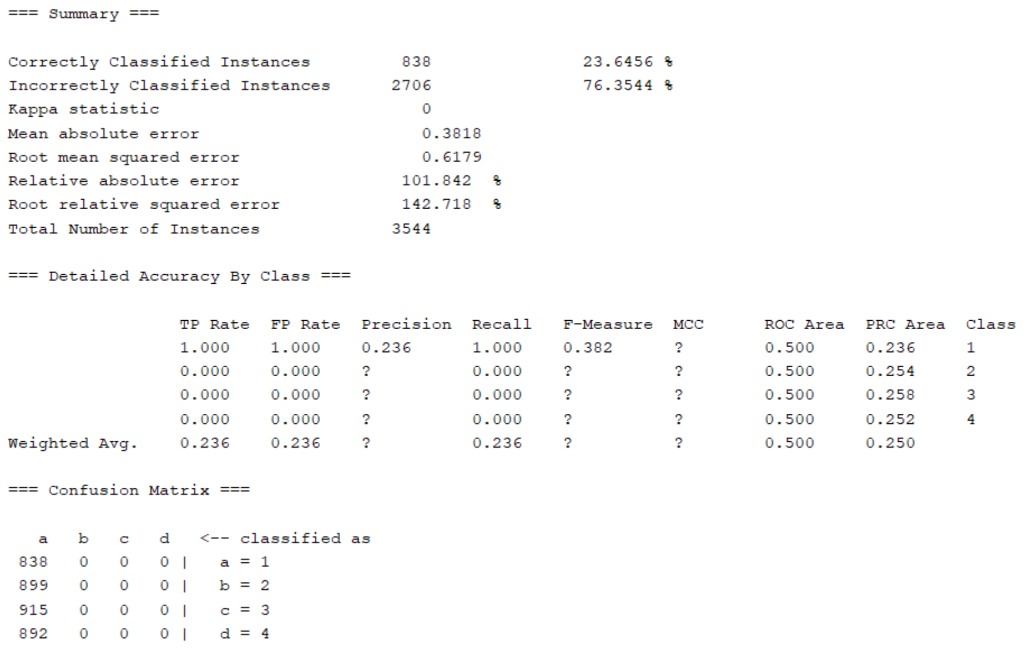
6. IMPLEMENTATION

After cleaning the dataset and fixing missing values, lies the question of extracting information from them. The dataset here is unlabelled and thus unsupervised learning was our first candidate.
First model used to cluster the data was simple k-means. But there was an issue as we didn't know how many clusters were optimal for this dataset, and should we use all 5 attributes that determine the weather over many buoys to cluster. The elbow method was used to find the optimal value of k over the features. The plot of k vs sum of squared



```
=== Summary ===

Correctly Classified Instances        1009               28.4707 %
Incorrectly Classified Instances      2535               71.5293 %
Kappa statistic                          0.0616
Mean absolute error                      0.3576
Root mean squared error                  0.598
Relative absolute error                 95.4064 %
Root relative squared error            138.135  %
Total Number of Instances             3544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.909    0.254      1.000    0.405      0.152    0.546     0.254     1
                 0.057    0.014    0.580      0.057    0.103      0.120    0.521     0.272     2
                 0.078    0.015    0.645      0.078    0.139      0.158    0.531     0.288     3
                 0.055    0.000    1.000      0.055    0.104      0.204    0.527     0.293     4
Weighted Avg.    0.285    0.222    0.625      0.285    0.184      0.159    0.531     0.277

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 838   0   0   0 |   a = 1
 848  51   0   0 |   b = 2
 825  19  71   0 |   c = 3
 786  18  39  49 |   d = 4
```

Figure 3: Classification Rule based using PRISM with target class month



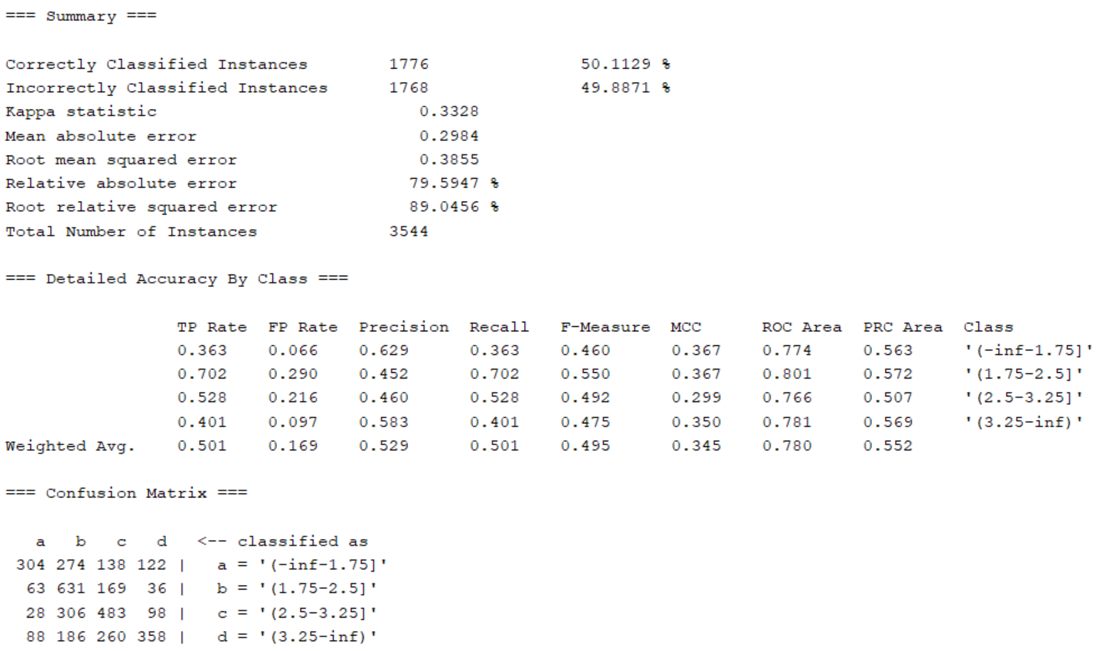
```
=== Summary ===

Correctly Classified Instances         838               23.6456 %
Incorrectly Classified Instances      2706               76.3544 %
Kappa statistic                          0
Mean absolute error                      0.3818
Root mean squared error                  0.6179
Relative absolute error                101.842  %
Root relative squared error            142.718  %
Total Number of Instances             3544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.236      1.000    0.382      ?        0.500     0.236     1
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.254     2
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.258     3
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.252     4
Weighted Avg.    0.236    0.236    ?          0.236    ?          ?        0.500     0.250

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 838   0   0   0 |   a = 1
 899   0   0   0 |   b = 2
 915   0   0   0 |   c = 3
 892   0   0   0 |   d = 4
```

Figure 4: PRISM with target class month and only SS Temp and Air Temp

error was used with k ranging from 1 to 15.

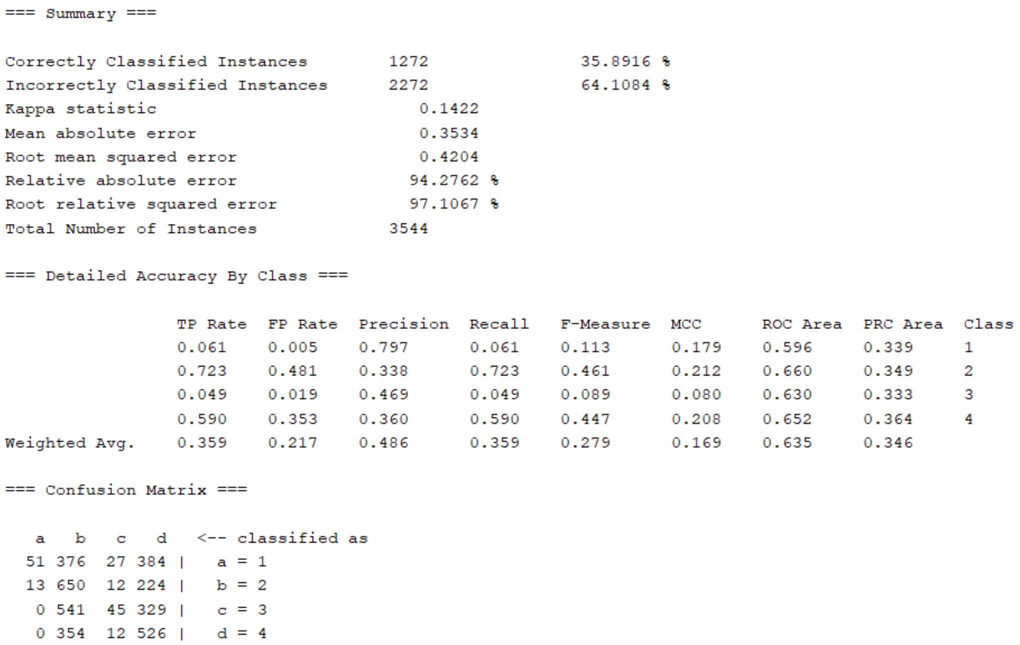
As mentioned before, the decision of using all or some attributes was the question. So, we tried k means over 5 attributes first and then used 2 attributes later. 5 attributes resulted in more clusters than two attributes. The result section talks about the feature selection for clustering. In short, 5 features meant 5 dimensions and the curse of dimensionality caused these dataset to spread out, cause bad cluster quality. Then two features were chosen that had the highest covariance between them, this reduced the features from 5 to 2 and improved the cluster quality. When all 5 attributes were used, we got an optimal k value between 4 and 5 with bad cluster quality, but when 2 attributes were used, we got an optimal k value between 2 and 3 with better cluster quality.

Now that we have the clusters, we need to find the repre-



```
=== Summary ===

Correctly Classified Instances        1776               50.1129 %
Incorrectly Classified Instances      1768               49.8871 %
Kappa statistic                          0.3328
Mean absolute error                      0.2984
Root mean squared error                  0.3855
Relative absolute error                 79.5947 %
Root relative squared error             89.0456 %
Total Number of Instances             3544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.363    0.066    0.629      0.363    0.460      0.367    0.774     0.563     '(-inf-1.75]'
                 0.702    0.290    0.452      0.702    0.550      0.367    0.801     0.572     '(1.75-2.5]'
                 0.528    0.216    0.460      0.528    0.492      0.299    0.766     0.507     '(2.5-3.25]'
                 0.401    0.097    0.583      0.401    0.475      0.350    0.781     0.569     '(3.25-inf)'
Weighted Avg.    0.501    0.169    0.529      0.501    0.495      0.345    0.780     0.552

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 304 274 138 122 |   a = '(-inf-1.75]'
  63 631 169  36 |   b = '(1.75-2.5]'
  28 306 483  98 |   c = '(2.5-3.25]'
  88 186 260 358 |   d = '(3.25-inf)'
```

Figure 5: RandomForest with target class as month



```
=== Summary ===

Correctly Classified Instances        1272               35.8916 %
Incorrectly Classified Instances      2272               64.1084 %
Kappa statistic                          0.1422
Mean absolute error                      0.3534
Root mean squared error                  0.4204
Relative absolute error                 94.2762 %
Root relative squared error             97.1067 %
Total Number of Instances             3544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.061    0.005    0.797      0.061    0.113      0.179    0.596     0.339     1
                 0.723    0.481    0.338      0.723    0.461      0.212    0.660     0.349     2
                 0.049    0.019    0.469      0.049    0.089      0.080    0.630     0.333     3
                 0.590    0.353    0.360      0.590    0.447      0.208    0.652     0.364     4
Weighted Avg.    0.359    0.217    0.486      0.359    0.279      0.169    0.635     0.346

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  51 376  27 384 |   a = 1
  13 650  12 224 |   b = 2
   0 541  45 329 |   c = 3
   0 354  12 526 |   d = 4
```

Figure 6: RandomForest with target class as month, ss temp and air temp as attributes

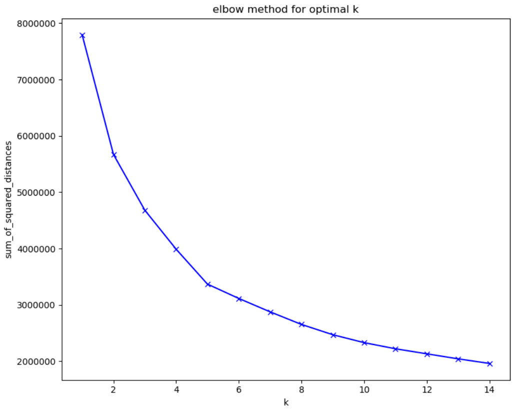
sentation. El-nino dataset helped here in finding the representation of the cluster. This dataset represented high winds and possibility of storm or other climactic warnings and so we decided to use them to predict which of the cluster they would belong to. The maximum number of points in the cluster would represent that the cluster can be used for prediction of storm. For two attributes, we used k=2 as optimal value after using elbow method. When El-Nino was used for prediction, more samples went into cluster 1 than cluster 2. The two clusters from the TAO data have their own probability distribution curve, so probability of data being in those clusters can be found.

We used the same concept with DBscan but the results were not satisfactory, and so we decided on using k-means for modelling.

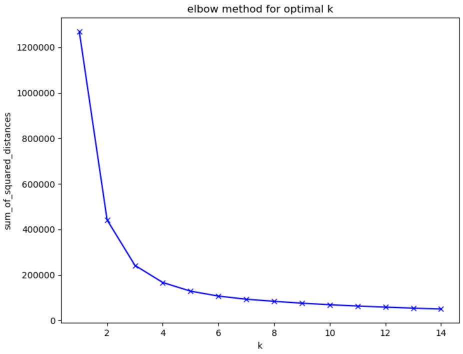
7. RESULTS

The clustering process resulted in the following measurements. First elbow method was used to find the optimal k value for 5 attributes over TAO dataset. k = 1 to 15 was used. k vs sum of squared errors was plotted.

All 5 attributes used



2 attributes used



For 5 attributes, elbow was found near k = 4 and 5 and For 2 attributes, elbow was found near k = 2 and 3

The table shows algorithm used, k value found and the silhouette score.

Table 1: Different Models

Model	Clusters 'k'	Silhouette Score
k means all 5 attr	4	0.259030788
k means 2 attributes	2	0.669598072
DBScan	4	-0.29821332

8. CONCLUSION

After working with clustering and classification on the dataset, and using different models again on each type, it was evident that for unlabelled data like the one used here, clustering seemed to be an appropriate choice. K means here gave promising results with a silhouette score of 0.669 which is high enough to consider the cluster to be of a good quality.

9. FUTURE WORK

For improving the prediction model, it is important to find the relations and patterns in which attributes interact with each other. Thus effort needs to be spent on distance calculation formula. Also, there other data mining algorithms that can perform on sparse datasets. We can also explore concepts from neural networks like generative adversarial networks and apply them to our dataset. This will help us find the hidden patterns as the attributes are probably not related with each other at polynomial scale. Thus in future, various algorithms need to be tested on this dataset for us to capture the hidden pattern.

10. REFERENCES

[1] The 1982-83 el nino. data retrieved from `https://www.fcst-office.com/HardRock/Meteo241/El%20Nino%201982-1983/ProjectThree.html`.

[2] Dbscan. `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`.

[3] El nino dataset. data retrieved from `https://archive.ics.uci.edu/ml/datasets/El+Nino`.

[4] Global tropical moored buoy array. data retrieved from `https://www.pmel.noaa.gov/gtmba/mission`.

[5] K means. `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`.

[6] National weather services. data retrieved from `https://www.weather.gov/mhx/ensowhat`.