Monday, August 3rd , 2015

SURA HUNTERS

TEAM MEMBERS

- Arunkumar Bagavathi
- Guoquing Yu
- Nachiket Doke
- Naveen Kumar Ananthu Langaram
- Raja Kamaraj
- Rakesh Balan Lingakumar

OUTLINE

- Our Approach
- Master Data set
- Data pre processing
  - Data transformation
  - Feature selection
- Assumptions
- Algorithms
- Findings
- Future Work
- Q & A

OUR APPROACH

- Quantify potential causes leading to increased shark attacks along North Carolina sea shore.
- Optimize data - better results.
- Machine learning algorithms - understand correlation among environmental factors.

MASTER DATA SET

- Six attributes:

1. Date
2. Land temperature
3. Sea surface temperature
4. Sea turtle nestlings
5. Salinity
6. Moon phases
7. Shark Attack - Class instance

DATA PREPROCESSING

Assumptions:

- Data set consists of Summer months from April to September.
- Negative class instances - to find out cases in which sharks did not attack.
- Extra class value ‘Maybe’ on some days in secondary data.

Data cleaning:

- Land temperature attribute - Missing value
- Salinity and Sea surface temperature - Numerical errors

DATA TRANSFORMATION

- Sea Surface Temperature and Salinity attributes - Weekly data - Aggregated using VBA Script

- Numerical attributes with high precision fractional data transformed into absolute values
- Achieved using *numericTransform* filter

- Discretization - Numeric attributes to Nominal attributes (5 discrete nominal values)

- UnderSampling using *SpreadSubSample* filter

FEATURE SELECTION

- WEKA process - Filtering method(InfoGainAttribute Evaluator and Ranker search method).
- Ranks the attributes by information gain with restect to the class.
- In our case, ‘Moon Phase’ is in lowest rank and is eliminated.

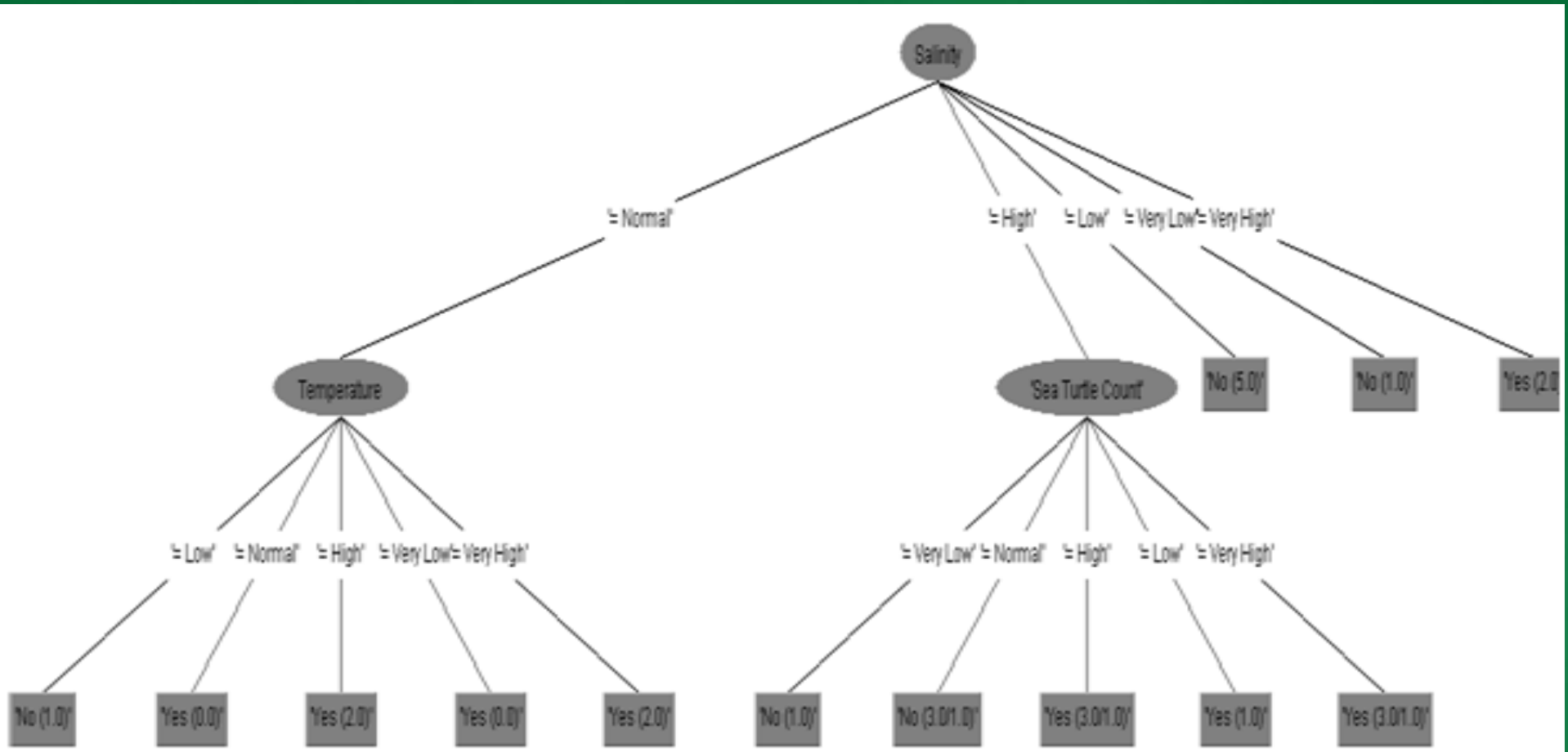
ALGORITHMS

1. OneR
2. NaiveBayes
3. IBk
4. J48
5. ClassificationViaClustering
6. Apriori

RESULT

Algorithm	Cross Validation (%)	Percentage Split (%)
One R(minBucketSize = 6)	66.67	75
Naïve Bayes	79.17	87.5
IBk (kNN = 5)	75	87.5
J 48 (confidenceFactor = 0.75)	62.5	75
ClassificationviaClustering (SimplekMeans	54.17	75
ClassificationviaClustering (EM)	70.83	62.5

J48 DECISION TREE



ASSOCIATION RULE

Confidence factor ranging from 0.6 to 0.67

Sea Turtle Count=Very High Salinity=High 3 ==> Temperature=Very High Water Temperature=Very High 2
Water Temperature=Normal Sea Turtle Count=Very Low Shark Attack=No 3 ==> Salinity=Low 2
Water Temperature=Normal Shark Attack=No 3 ==> Sea Turtle Count=Very Low Salinity=Low 2
Water Temperature=Normal Sea Turtle Count=Very Low 3 ==> Salinity=Low Shark Attack=No 2
Water Temperature=Normal 3 ==> Sea Turtle Count=Very Low Salinity=Low Shark Attack=No 2
Water Temperature=Very High Sea Turtle Count=Very Low Shark Attack=Yes 3 ==> Salinity=Very High 2
Sea Turtle Count=Very Low Shark Attack=Yes 3 ==> Water Temperature=Very High Salinity=Very High 2
Water Temperature=Very High Sea Turtle Count=Very High Shark Attack=Yes 3 ==> Salinity=High 2
Water Temperature=Very High Sea Turtle Count=Very High Salinity=High 3 ==> Shark Attack=Yes 2
Sea Turtle Count=Very High Shark Attack=Yes 3 ==> Water Temperature=Very High Salinity=High 2
Sea Turtle Count=Very High Salinity=High 3 ==> Water Temperature=Very High Shark Attack=Yes 2
Water Temperature=Very High Salinity=High 8 ==> Shark Attack=Yes 5
Sea Turtle Count=High 5 ==> Temperature=Very High 3
Sea Turtle Count=High 5 ==> Water Temperature=Very High 3
Sea Turtle Count=High 5 ==> Salinity=High 3
Sea Turtle Count=High 5 ==> Shark Attack=Yes 3
Water Temperature=Very High Sea Turtle Count=Very Low 5 ==> Temperature=High 3
Water Temperature=Very High Sea Turtle Count=Very Low 5 ==> Shark Attack=Yes 3
Salinity=High Shark Attack=No 5 ==> Water Temperature=Very High 3
Water Temperature=Very High Shark Attack=No 5 ==> Salinity=High 3

SCENARIO 2

In this scenario, we introduce a new instance ‘May be’ in the dataset. The purpose of introducing a new class instance into the data set improved the overall prediction/accuracy based on the below factors,

- The ‘May be’ instance is used in our data in order to support the ‘Yes’ class instance and thus we have assumed this instance defines a likely chance of shark attack occurrence.
- We made the assumptions based on the obtained domain knowledge and basic habitats of sharks.

Logic used to determine “may be” instance

If (Shark Attack is Yes) then

Check for next shark attack

If (Shark attack within 15 days)

Apply Instance as ‘May be’

Else

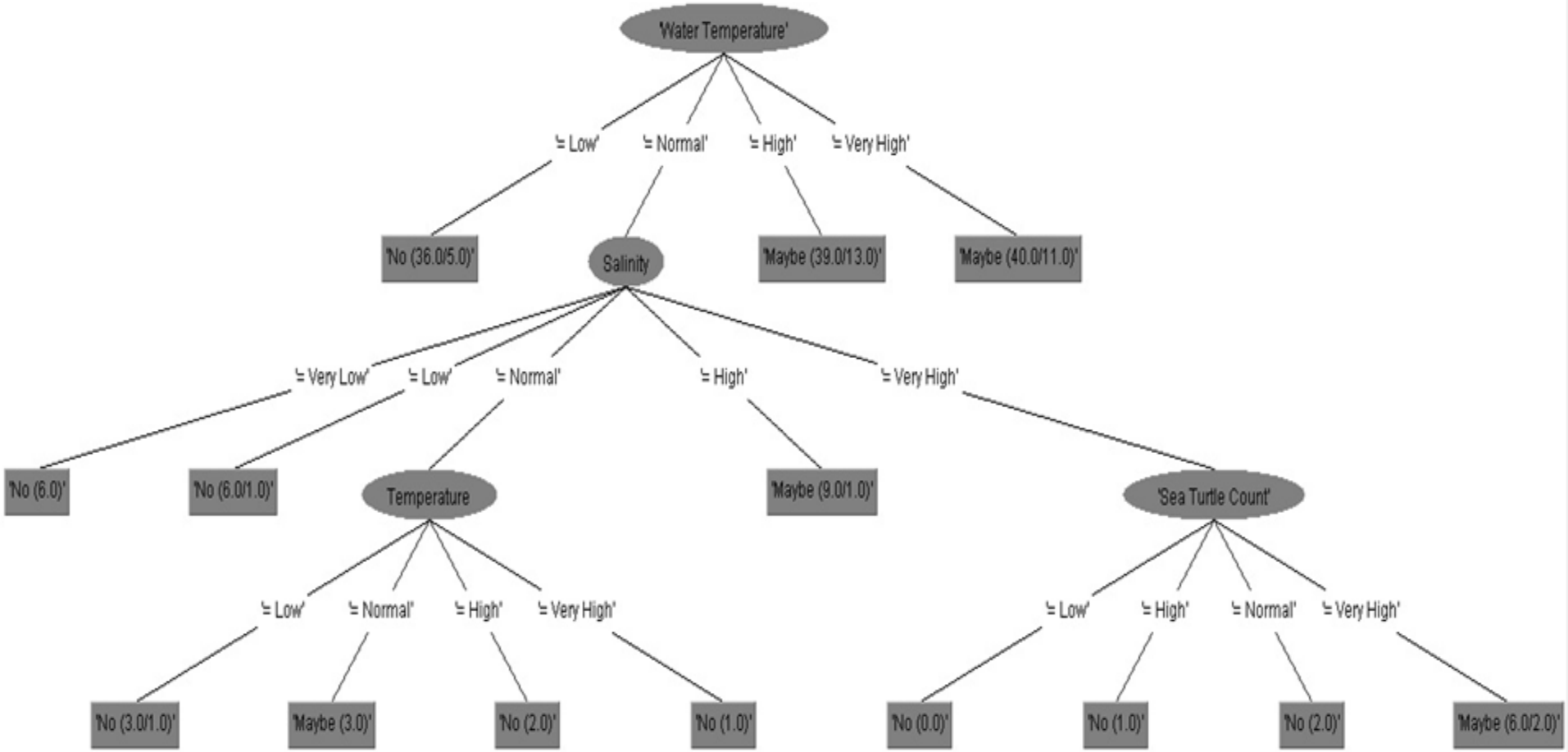
Check for common attributes where shark attack likely to happen

Apply Instance as ‘May be’

Results from Scenario 2

Algorithm	Cross Validation (%)	Percentage Split (%)
One R(minBucketSize = 6)	70.1299	65.3846
Naïve Bayes	70.7792	71.1538
IBk (kNN = 5)	69.4805	67.3077
J 48 (confidenceFactor = 0.75)	75.3247	82.6923
ClassificationviaClustering (SimplekMeans)	47.4026	42.3077
ClassificationviaClustering (EM)	59.0909	44.2308

J48 DECISION TREE for Scenario 2



ASSOCIATION RULE for Scenario 2

Best rules found:

1. Water Temperature=Low 36 ==> Shark Attack=No 31 conf:(0.86)
2. Salinity=Very Low 20 ==> Shark Attack=No 17 conf:(0.85)
3. Salinity=Very High 33 ==> Shark Attack=Maybe 25 conf:(0.76)
4. Temperature=Low 24 ==> Shark Attack=No 18 conf:(0.75)
5. Water Temperature=Very High 40 ==> Shark Attack=Maybe 29 conf:(0.73)
6. Water Temperature=High 39 ==> Shark Attack=Maybe 26 conf:(0.67)
7. Temperature=Very High 45 ==> Shark Attack=Maybe 28 conf:(0.62)
8. Sea Turtle Count=Very High 57 ==> Shark Attack=Maybe 35 conf:(0.61)
9. Salinity=Low 30 ==> Shark Attack=No 18 conf:(0.6)
10. Salinity=High 36 ==> Shark Attack=Maybe 21 conf:(0.58)

Usual Suspects

According to NBC news, sharks that could be responsible for NC attacks are,

- BlackNose Shark
- BlackTip Shark
- Spinner Shark
- Tiger Shark
- Bull Shark

The first three sharks generally don't attack humans and their attack may due to high salinity content in seawater.

The last two sharks are aggressive and tends to attack anything that moves in their territory. They tend to move around the shore for their favourite meal "sea turtle".

FUTURE STUDY

1. Collecting data sets along the east coast of United states and evaluate the findings
2. Update the data set with environmental factors such as phytoplankton counts, chlorophyll count and sea surface currents to evaluate and improve our correlation mapping.
3. Optimize the parameter setting in WEKA for generating better results.
4. Focus based analysis should be made on the sharks which attacked.

REFERENCES

Shark attacks - NC	http://www.sharkattackdata.com/gsaf/place/united_states_of_america/north_carolina
Land temperature	http://w2.weather.gov/climate/xmacis.php?wfo=mhx
Sea surface temperature	https://www.nodc.noaa.gov/cwtg/all_meanT.html
Sea turtle nestlings	http://www.seaturtle.org/nestdb/index.shtml?view=1&year=2015
Salinity	http://carolinasrcoos.org/queryStation.php

UNC CHARLOTTE