An Investigation into the Correlation between Environmental Factors and Levels of Snail Density

Authors: Terence Fusco, Yaxin Bi, Chris Nugent and Shengli Wu.

Introduction

This report presents an investigation into the relationship between environmental factors and snail density levels in the Dongting Lake region of China. The work involves preprocessing of remote-sensing data, significant analysis for further processing to build a classification system, and statistical methods, and the correlation analysis between the environmental factors and the density of snails. With the support of the Chinese partner, various comparative studies have been carried out: the results show those environmental factors that have the considerable influence on the density and distribution of the snail vector. From the above process, there is sufficient scope for future application of results on alternative databases. The general proposition of the findings reveals that NDMI and NDVI moisture and vegetation indexes are intrinsically linked with the high levels of snail density and distribution.

Objectives

1. Find the most meaningful relationships between each individual environmental factor and the density and distribution of snails in that area
2. Extract information which we can provide to local authorities and municipalities with individuals with a risk of exposure to vector-borne diseases
3. To develop a new machine learning approach in order to build a better classification and prediction model for future disease outbreaks
4. Make highly accurate oral density predictions based on environmental factors present

Methods

1. Plotting paired attributes with Snail Density – To assess relationship correlations with Snail Density levels
   - We can see from the scatter-plot diagram the relationship between the Normalised Difference Moisture Index (NDMI) and the Normalised Difference Vegetation Index (NDVI) in relation to snail density (snails)

2. Information Gain – To assess the attribute values in relation to snail density. Information gain attribute ranker was used and documented in a table for each dataset. The results show a relatively consistent approach with each having similar rankings for each of the attributes.

   - Information gain is a way of ranking attributes (Si) when given the class information (C), it makes calculations based on the probability of the class given the attribute (p(C|Si)). By this method we can make a better analysis of the attributes in the way to relation to the class.

3. Correlation Analysis – This method was applied to the combination of each of the attributes with the snail density temporarily. In terms of relationships prediction of several axes, the Tasseled Cap features were used to show the association between two attributes and the snail density for that year.

   - This method allows us to provide analysis on each individual relationship between the attributes and snail density as well as overall yearly data trends

4. Cumulative Training Approach (CTA) – To enrich the dataset by using the training potential in a limited set of dataset
   - The CTA enables us to make the most of our limited sized dataset by combining multiple years training data with alternative years testing data. This gives us a new insight into the similarities between the data from each year and provides us with useful information on the benefits of using this method for enriched training and testing on future data.

5. Replace missing values – To get a value for each of the missing values in the dataset, it was decided to use the mean and median values from the dataset. The replace missing values filter from Weka was used with the dataset and then a sample was taken. An exception was made for data with missing values. Weka is able to verify the accuracy of the results which are displayed in the table and chart in this paper.

   - The Replace missing values method has enabled us to handle partially complete datasets which will be an issue for any future testing due to the nature of the data which is collected by different agencies and cities. If we can provide a sufficient solution to this problem, we can apply it with confidence for future incomplete data.

Information Gain Attribute Ranker

Pearson’s Correlation coefficient

Pearson’s correlation coefficient is a measure that shows the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to +1, where:

- A value of 1 indicates a perfect positive correlation (as one variable increases, the other variable increases).
- A value of -1 indicates a perfect negative correlation (as one variable increases, the other variable decreases).
- A value of 0 indicates no linear correlation.

\[
p(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}
\]

where:

- \( \text{cov}(x, y) \) is the covariance between variables x and y.
- \( \sigma_x \) is the standard deviation of x.
- \( \sigma_y \) is the standard deviation of y.

Handling Missing Values

Proposed semi-supervised model

The diagram above provide a depiction of our proposed semi-supervised classification model. The main processes in the model diagram include:

1. Labelling large unlabeled data using smaller labeled data as seeds for labeling
2. Building a prediction model based on the labeled dataset
3. Providing accurate and consistent classification of snail density for use in prediction of disease likelihood

By investigating this area, we can assess the value of various classifiers in relation to snail density and distribution and the environment factors. Ensemble learning methods are ways of combining classifiers to make use of the best performance qualities of each ensemble. Further to this research, the subsequent plan is to begin experimenting with cluster ensemble classification in order to investigate the potential of working with feature consensus and subsets of the data on classification ensemble[3]. This is an area which is currently being implemented that can be applied to different scenarios.

Summary

From the results of preliminary experiments on the dataset, we can see that the stacking and Naive Bayes classification accuracy is comparatively relative but with higher accuracy, the stacking classifier edge Naive Bayes as in the case with years 2000 and 2001. The stacking classifier can now be considered for future work based on accuracy result due to the objective for higher classification accuracy. Naive Bayes generally can only compete with consistency when lower levels of accuracy are needed. The stacking classifier is also a robust classifier as it can use many alternatives such as SVM[1] and ensemble learning classifiers in practice.

We can use from the information gain attributes rankings from each year, that the four highest ranking features from each year are the NDMI, TC, NDVI and NDVI with the exception of 2009. This work would later be the focused upon attributes of TC, NDVI and TC B could be further research on a subset in the data that may carry significant influence on the snail density and distribution outcomes [3]. It is possible to deduce from a subset of the data, a higher degree of classification accuracy than that of the entire dataset in each case, then we may be able to select the most pertinent attributes with which to make predictions while also reducing the amount of busy data and resources that exist.

Looking at the Support Vector Regression results in the table, we can see that the SVM model had consistently the highest error rate of any year. This is the same for the SVR model and logistic regression which shows that this value replacement factor of raw data, this is not the case for the Support Vector Regression systems [1] and we cannot use the data we have been using for the stacking classifier.

References