A Cumulative Training Approach to Schistosomiasis Vector Density Prediction

Terence Fusco(✉) and Yaxin Bi

School of Computing and Mathematics, Ulster University, Newtownabbey, UK
Fusco-T@email.ulster.ac.uk, bi.y@ulster.ac.uk

Abstract. The purpose of this paper is to propose a framework of building classification models to deal with the problem in predicting Schistosomiasis vector density. We aim to resolve this problem using remotely sensed satellite image extraction of environment feature values, in conjunction with data mining and machine learning approaches. In this paper we assert that there exists an intrinsic link between the density and distribution of the Schistosomiasis disease vector and the rate of infection of the disease in any given community; it is this link that the paper is focused to investigate. Using machine learning techniques, we want to accumulate the most significant amount of data possible to help with training the machine to classify snail density (SD) levels. We propose to use a novel cumulative training approach (CTA) as a way of increasing the accuracy when building our classification and prediction model.

1 Introduction

The resurgence of epidemic disease breakouts in regions of Asia and South America in the past decade has given local governments and health organisations cause for much concern. The devastating impact that these diseases can have on many aspects of human, cattle and crop life incurs huge financial and social cost. This rationale makes research into the prevention and preparation for future outbreaks, a problem that requires immediate attention and one that is crucial to supporting the locally affected municipalities [1]. The epidemic disease Schistosomiasis is detrimental to many sections of society in China. Schistosomiasis is the second most widely affected disease in the world as stated by the World Health Organisation [2]. It is a disease, which is transmitted through water infected by parasites known as Schistosomes. The intermediate host of the disease is the Oncomelania Hupensis snail. Humans are affected mainly through freshwater used for washing clothes and household items as well as through infected crops and cattle. The affect it can have on many areas of human, cattle, crop life both in terms of health and financially is a valid cause for concern [1]. To combat Schistosomiasis can be very difficult due to the fact that there is no vaccine available against the disease and therefore it can only be treated once the patient has been infected. Currently, the most effective way of dealing with the disease is by

L. Iliadis and I. Maglogiannis (Eds.): AIAI 2016, IFIP AICT 475, pp. 3–13, 2016.
DOI: 10.1007/978-3-319-44944-9_1

trying to establish areas that are of high risk of the disease and putting in place preventative measures such as chemical treatment to specific freshwater areas [3] in order that the disease is addressed before the vectors multiply or increase in density and distribution. An alternative solution is to plant poplar trees, which would disturb the natural vegetation and moisture factors that encourage snail life and breeding habitat [4]. Whichever method is applied to at-risk areas will have a time and financial cost incurred therefore the concerned municipalities require the most informed data available before acting and addressing the area in question. The local governments will also need to prepare those areas for any panic or influx of patients that may occur.

The environment features present in Schistosomiasis areas of interest can be shown to be intrinsically linked to the disease infection rates [5]. By using data mining methods we can assess the corollary relations between the environment feature values and the SD and distribution values. We aim to identify the environment conditions which make the Oncomelania Hupensis snail most suitable for transmission of the Schistosomiasis disease. We know that for the Oncomelania Hupensis reproduction and for life to flourish, it requires specific environment conditions. We also know the snails will not survive in strong currents and that during early years in their lives the Oncomelania Hupensis snail will live only in water. Once they are adults they then must move from the water usually to moist soil above the water line as the snail activity increases with soil moisture and that the optimum temperature for breeding is around 20 °C [6]. The Oncomelania Hupensis snail flourishes and breeds particularly well in areas with high levels specific environment features such as soil moisture (NDMI) and vegetation (NDVI) therefore we can deduce that areas which meet these specific environment conditions have a greater likelihood of high snail vector density.

By analyzing and assessing this information we can achieve greater success from our classification accuracy. With the implementation of this research approach we can make the most informed prediction on which to base information to provide to those concerned. We believe that the most promising approach to detect high-risk areas of disease outbreak is to use vector density classification techniques based on environmental features that exist in each area of interest.

Using our proposed Cumulative Training Approach (CTA), we can enhance the training potential of our limited dataset. This will help to provide a larger pool of relevant training data which we hope will increase the classification accuracy during the testing process. The process involved uses the data from a combination of collective years' data as a training set to train the machine for classifying SD based on the environment information given. Particularly the CTA also involves the pre-processing of segmenting the SD into the three or five point categories, handling of missing values, environment feature selection and correlation analysis between environment features and attributes.

This paper provides the description, rationale and results of preliminary experiments that examine the correlation and influence levels between environmental features and SD present in the Dongting Lake area of China. This lake represents a very relevant study area with which to examine the moisture and vegetation levels required for snail life to flourish. The datasets used in this paper were derived from remotely sensed image extraction information together with manually collected field survey data provided by our Chinese project partners at Academy of Opto-Electronics the and the European Space Agency (ESA). The datasets have been analysed quantitatively and results are illustrated with this paper. The aim of these studies is to discover if there exists strong correlation between individual or component environment features and the Schistosomiasis disease vector (Oncomelania Hupensis snail) density and distribution. If we can identify this, then we can make future SD classifications using our prediction models based on previously collected datasets. The resulting prediction models will be capable of making informed assumptions on future SD levels and therefore provide likelihood of outbreaks of the disease occurring based on environmental feature values on a larger scale and with greater efficiency than is currently available.

The CTA proposed in this paper is a framework we use to enhance the training potential of real-world sparse datasets. This approach is conducted using a range of pre-processing methods together with attribute ranking and data analysis, the results of which we take into consideration when building our classification and prediction models to determine the density and distribution of epidemic disease vectors. We aim to enrich our training set by investigating the various methods discussed in this paper to deduce if they can have a positive effect when applied for classification of SD in terms of accuracy performance. This includes using combined years of data instances for training against alternative testing sets. We aim to take into account and apply the optimal test conditions as a training paradigm to discover whether those criteria perform better than standard datasets for classification purposes.

2 Experiment Data

The datasets used in this report are derivative of remotely sensed satellite images ranging from between 2003–2009 in the Dongting Lake area of China. The images were processed and feature extraction was carried out by our Chinese partners to provide values for the environmental features present for each year. While we can access vast amounts of data from satellite imaging, the primary field survey data of which we can be sure is a much more time consuming process so this is why we have such a limited dataset in terms of instances. The number of common features from each year was seven with the collective number of instances being 180. While the dataset is relatively small in data mining terms, it provides a basis on which to form initial opinions and observations as to which attributes or combination of attributes have the strongest influence on SD levels. When we deduce which feature subsets are most influential on the SD levels, we can make assertions on future SD classification and therefore provide important information to those concerned for preventative measures to be put into place.

During initial assessment of the dataset, we looked at how we would categorize the SD values in terms of whether the raw value provided would constitute the label of high SD. To this end, the data was preprocessed by normalizing the values in order to gain arbitrary values into a predefined range which could then be labelled in terms of density level. We subsequently assessed whether we could achieve better classification success by using a 3 or 5 point scale of SD as in Low, Medium or High as opposed to Very Low, Low, Medium, High and Very High. We must categorize the density level in this way for classification purposes otherwise we will be restricted to using either statistical or regression models. The data was initially normalised to achieve a range from 0 to 1 then discretised into the 3 or 5 point scale with even distribution. To discretise the data, data binning was used. This method takes a set of continuous values and turns them into a set of bins which are nominal values. The number of bins was set to five for 'very low', 'low', 'medium', 'high' and 'very high'; this resulted in five data intervals which the data was then split into. It was also used for three bins, representing 'low', 'medium' and 'high'.

In addition, we used two regressive methods on our data to make initial assessment on how well the data fits and therefore how well each year fits for classification purposes. While using the linear regression and support vector regression on our unprocessed SD data, we can assess the accuracy of each year of data when predicting new instances of SD. With linear regression we assume environment features to be independent in a dataset; in this case the environment factors are in relation to the dependent variable which is the SD value.

These regressive methods do not provide specific classification percentage accuracy results as the SD value has not been pre-processed or classified to a selected scale that can be used for prediction. We can instead use the coefficient of determination calculation to give the R value which tells us how good of a fit the data is that we are experimenting with. We can assess the results of the coefficient of determination with results ranging from 0 to 1. The closer to 1 the result is, the better the fit therefore the higher likelihood of predicting a new instance of SD.

Equation 1 involves taking the average of the entire SD actual values for each year, then subtracting the average value from each individual actual SD value to the power of two and the same with each predicted value from that year. We then take the sum of results from each instance of predicted and actual values and divide the total value from the predicted SD calculation by the total actual value calculation with the value ranging between 0 and 1 with 0 being the least well-fitting data and 1 being the best fitting data.

We can see from Table 1b that the best fitting data is from 2008 training and testing data using linear regression. It scored 0.8 which is the closest result to 1 making it the most promising data combination for classifying future instances of SD. We can also see that in 2007, both the linear regression and support vector regression classifiers performed well with similar performance which can also be an indicator of potentially generating good classification models.

Once we had made initial assessment of the datasets, they were preprocessed by normalizing and then discretizing the SD information from each year. This enabled us to have more options for using different algorithms for classification

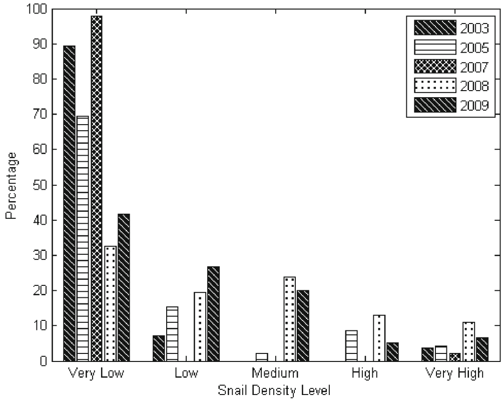


Fig. 1. Classification of SD over Time

Table 1. Statistics for yearly SD Categories and R^2 Values

Year	AVG SD	Normalised	Category
2003	0.727	0.107	V. Low
2005	0.879	0.129	V. Low
2007	2.633	0.388	Low
2008	1.396	0.206	Low
2009	1.056	0.156	V. Low
Collective	1.014	0.149	V. Low

(a) Average SD Values

	2003	2005	2007	2008	2009
LR	0.325	0.590	0.734	0.808	0.699
SVR	0.052	0.221	0.732	0.691	0.506

(b) R^2 Values for Linear Reg. and Support Vector Reg.

purposes. The SD data was separated at the beginning into 5 categories and the results are recorded in Fig. 1 and Table 1a. They show on average that density and distribution of the Oncomelania Hupensis snail during the time period from 2003–2009 were predominantly very low and low. This is what we would expect to see once the data has been preprocessed but it does not provide the entire picture so we will have to explore the dataset further and assess the relative SD levels and in conjunction with environmental features.

3 Methods

In earth observation research, weather conditions directly affect the quality of satellite imagery which causes some values of environment variables to be missing in the set and discontinuity in terms of fully recorded data relationships. These partially complete datasets are caused by anomalies in the remotely sensed image extraction process and by issues such as weather clarity from satellite imagery.

One of the major issues faced when using the data provided was that specific data particularly from 2007 was only partially complete. This problem correlates directly to the weather conditions present at the time of acquiring data from the satellites therefore we are interested in providing a resolution that can be applied to any future incomplete datasets provided by satellite images. This issue highlighted the need for an approach capable of imputation of the values that were incomplete from the dataset in order to be able to use the 2007 dataset and any other incomplete data for future temporal assessment of SD levels.

The rationale behind this imputation process is to find a solution for replacement of partially complete data that could potentially be scalable for much larger datasets with a variety of different features. The process of removing known values from our dataset then providing replacement using the following methods is documented below, where V represents the feature value of an instance.

- The Weka replace missing value filter replaces missing values with the mean and modal values from the remaining set for data imputation.
- The Single Pre-Succession Method uses the previous and following values to replace the missing value.
$$v_i = \frac{v_{i-1} + v_{i+1}}{2} \tag{1}$$
- The Mean Single Pre-Succession Method uses the previous and following values to replace the missing value together with the mean of the entire set.
$$v_i = \frac{v_{i-1} + v_{i+1} + \hat{v}}{3} \tag{2}$$
- The Double Pre-Succession Method uses the two previous and following values to replace the missing value.
$$v_i = \frac{v_{i-2} + v_{i-1} + v_{i+1} + v_{i+2}}{4} \tag{3}$$
- The Mean Double Pre-Succession Method uses the two previous and following values to replace the missing value together with the mean of the entire set.
$$v_i = \frac{v_{i-2} + v_{i-1} + v_{i+1} + v_{i+2} + \hat{v}}{5} \tag{4}$$

We can see from Table 2 that the most accurate performing method is the Mean Double PreSuccession method with an average percentage difference of 32.58 % while the lowest performance is of the PreSuccession method which has an average percentage difference of 333.36 % from the original value that was replaced. These results can now be analysed and used for future incomplete datasets to verify the accuracy of value replacement over more extensive datasets.

3.1 Feature Assessment

- We want to evaluate the dataset to discover the relevance of each attribute to SD levels individually and as subsets of features.
- To assess and rank the features of the data yearly to gain a deeper understanding of the value of the environment features to the data as a whole.

- Selection of an efficient, well performing method to handle replacement of missing values in the data as this is an ongoing issue with RS images that will be required for application in any future data that may be accessed.
- To distinguish the most effective category of SD to move forward with for future classification purposes.

3.2 Information Gain

To assess the attribute values in relation to SD, Information Gain attribute ranker was applied to the data and documented in a table for each years' data. Information Gain is a feature ranking approach that uses entropy to identify which feature in the dataset gains the most information relative to the class. This is beneficial when carrying out analysis of a dataset to extract the most influential features in relation to their corresponding SD values. It can be of significant value in order to identify any corollary inferences with regards environment features to SD levels. Once we identify which attributes have most significant influence on the SD value, then it can be established for future experiments that these specific attributes are closely connected to high levels of SD.

The results in Table 3 show relative consistency with each year having similar positions for each of the attributes. We can see certain attributes consistently trending such as the Normalised Difference Water Index (NDWI) and the Tasseled Cap Greenness (TC_G) which indicate that these attributes are of significant value in relation to SD of each of the particular years in the dataset. These results will now form an element of consideration for our CTA model.

The information Gain calculation used is shown in Eq. 5 where entropy (H) is given of the class (C) given the attribute (A) [7]. Entropy and information gain are intrinsically linked as the decrease in the entropy of the class is a direct

Table 2. Data imputation from 2007

Original	Weka	PreSucc. Method	Mean PreSucc.	Double PreSucc.	Mean Dbl PreSucc.
0.0348	0.228	0.251	0.169	0.012	0.061
0.128	0.201	0.100	0.254	0.123	0.069
−0.084	−0.084	0.236	0.052	0.401	0.064
0.024	0.201	0.097	0.106	0.027	0.050
−0.660	−0.471	0.026	−0.156	0.080	−0.083
0.242	0.228	0.386	0.214	0.103	0.072
−0.521	−0.622	0.387	−0.091	0.346	−0.062
0.410	0.657	−0.006	0.232	−0.141	0.112
0.400	0.344	0.025	0.123	−0.025	0.064
0.35545	0.657	0.007	0.236	−0.119	0.117
Avg.% Diff.	145.13 %	**333.36**%	226.81 %	132.04 %	**32.58**%

Table 3. Information Gain feature ranking

	2003	2005	2007	2008	2009	Collective
1	NDWI	NDWI	NDWI	TC_G	MNDWI	TC_G
2	TC_W	TC_W	TC_W	NDWI	NDVI	NDWI
3	TC_G	TC_G	TC_G	NDVI	NDWI	TC_W
4	NDMI	NDMI	NDMI	MNDWI	TC_G	NDMI
5	MNDWI	MNDWI	MNDWI	NDMI	NDMI	MNDWI
6	NDVI	NDVI	NDVI	TC_W	TC_W	NDVI
7	TC_B	TC_B	TC_B	TC_B	TC_B	TC_B

reflection of the added information about the class provided by the attribute and this is referred to as the information gain and therefore entropy is a pre-requisite for information gain to be calculated [8].

$$H(C|A) = -\sum_{a \in A} p(a) - \sum_{c \in C} p(c|a) \log_2 p(c|a) \tag{5}$$

Correlation analysis was applied to the combination of each of the attributes with the SD temporally. In terms of relationships, we used Pearson's r approach, which uses the covariates X and Y, this is then divided by the standard deviation of X and of Y to give a correlation value of each individual attribute and SD value. The results are shown in Fig. 2b and they indicate that data from 2008 is not in correlation with the alternate years as the trend lines show us. The combination of the SD and environmental attributes (X, Y) does not show correlate with the dataset from each year. The corollary relationship results between SD levels and environment features is an integral component of our CTA framework below for future classification of SD levels based on environment factors. As the environment feature values increase towards 1.0 in Table 4 it shows the impact factor that is present when compared with SD levels with TCG and NDVI from 2005 showing good correlation as opposed to NDWI in the same year.

$$P(x,y) = \frac{cov \quad (x,y)}{\sigma x \sigma y} \tag{6}$$

4 Cumulative Training Approach (CTA)

Given the limited amount of data and the pre-processing results, we consider how to construct prediction/classification models. A caveat to address with the SD classification is the fact that for years 2003, 2005 and 2007 we have 18/19 attributes partially complete labelled whereas with years 2008 and 2009 we have eight/nine attributes given to experiment with. The most beneficial approach to dealing with this issue is to use those attributes which are common to each

Table 4. Correlation analysis between SD vs. Features

	TCB	TCG	TCW	NDMI	NDVI	MNDWI	NDWI
2003	−0.069	0.352	0.352	0.101	0.359	−0.289	−0.339
2005	0.308	0.517	−0.301	0.332	0.519	−0.4999	−0.521
2007	0.287	0.192	−0.17	0.194	0.227	−0.07	−0.195
2008	−0.32	−0.132	0.163	0.194	−0.179	0.397	0.289
2009	0.208	0.333	−0.193	0.139	0.418	−0.462	−0.387

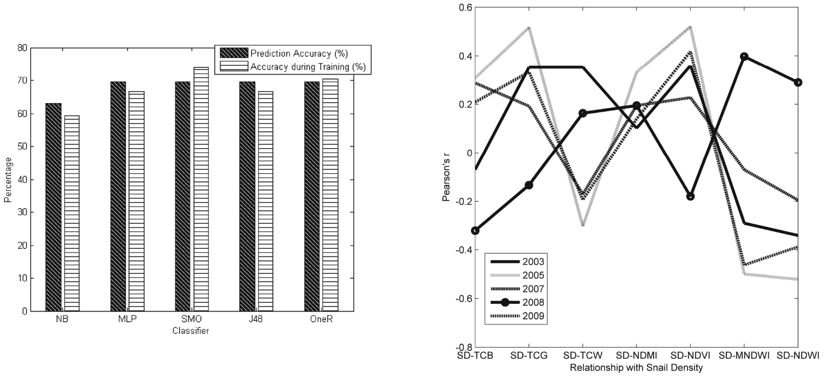
year in order to make a comparable dataset for training and testing purposes. It was decided to use the initial year's collected research information to build a training model, which is then used as a benchmark against future data for testing purposes. This approach will enable us to enrich the dataset with variable subsets of the data being used to discover temporal relationships within the dataset. The method was used tested with five classification methods to assess accuracy. We can see from Fig. 2a that year 2003 training data with 2005 testing yields highly accurate results as the accuracy during training and prediction accuracy are in close proximity to each other, this indicates good classification performance.

This method of training will make up another component of our proposed idea referred to throughout this paper as the CTA. By applying this proposed CTA, we are combining the most promising experiments and test results from a variety of relevant areas of our data pertaining to the classification of the Schistosomiasis disease vector. Using this data we can then build a model for application with any future spatio-temporal epidemic disease environment data, which is a different approach from the standard application of classification methods.

In addition carried out testing on three ensemble learning methods of Bagging, Boosting and Stacking as the ensemble methods have been shown to provide better classification accuracy than single classifiers [9]. Using these three ensemble methods we can get a varied range of results based on training model performance (Adaboost), equal sized training set sampling (Bagging) and combined classifier prediction (Stacking). Results were recorded in Table 5.

Table 5. CTA ensemble results

Training	Testing	Boosting		Bagging		Stacking	
		Train	Predict	Train	Predict	Train	Predict
2003/05	2009	0.740	0.483	0.753	0.5	0.712	0.467
2003/05/07	2009	0.530	0.467	0.504	0.467	0.556	0.467
2003/05/07/08	2009	0.558	0.483	0.509	0.533	0.491	0.467
2003/05	2008	0.740	0.478	0.753	0.348	0.712	0.326
2003/05/07	2008	0.530	0.326	0.504	0.348	0.556	0.326
2003/05	2007	0.740	0.295	0.753	0.318	0.712	0.295



(a) 2003Train - 2005Test CTA Data (b) Pearson's Correlation Co-Efficient

Fig. 2. CTA and Correlation figures

5 Conclusion

From the correlation analysis graph, we can see that each of the years data with the exception of 2008, follow together in a trend which shows that the correlation values of each combination of attributes together with SD, can be predictable which is of high value to this particular research area looking at future distribution and density predictions.

By handling and assessing missing value replacement in the data, we can identify the success of replacing these values based on the mean and mode of the existing data. These results can be applied to future RS data that will be accessed for research and experimentation. By testing effectiveness of replacement methods, we can identify confidence in future replacement of data.

All experiments and collective research to date have become part of the CTA for Schistosomiasis vector density and distribution prediction. This approach has provided us with a better understanding of our datasets and the classification results which it provides. In combining each aspect of the training process we have a greater understanding of the research area and we can apply this knowledge to future data obtained for classification and prediction purposes.

From the results to date, we can deduce that specific environmental attributes such as TC_G and NDWI have more influence on the SD and distribution than others. This information will be further analysed and implemented into a cluster ensemble algorithm for optimum accuracy classification for future work [10].

References

1. Ross, A.G.P., Sleigh, A.C., Li, Y., Davis, G.M., Williams, G.M., Jiang, Z., Feng, Z., Manus, D.: Schistosomiasis in the people's republic of china: prospects and challenges for the 21st century. Clin. Microbiol. Rev. **14**(2), 270–295 (2001)

2. WHO: Schistosomiasis (2015)
3. Ma, C., Dai, Q., Li, X., Liu, S.: The analysis of east dongting lake water change based on time series of remote sensing data. In: 2014 12th International Conference on Signal Processing (ICSP), Institute of Electrical & Electronics Engineers (IEEE), October 2014
4. Sun, Q., Zhang, J., Zhou, J., Wu, L., Shan, Q.: Effect of poplar forest on snail control in dongting lake area. In: 2009 3rd International Conference on Bioinformatics and Biomedical Engineering, Institute of Electrical & Electronics Engineers (IEEE), June 2009
5. Wu, J.Y., Zhou, Y.B., Li, L.H., Zheng, S.B., Liang, S., Coatsworth, A., Ren, G.H., Song, X.X., He, Z., Cai, B., You, J.B., Jiang, Q.W.: Identification of optimum scopes of environmental factors for snails using spatial analysis techniques in dongting lake region, china. Parasites Vectors **7**(1), 216 (2014)
6. Seto, E., Xu, B., Liang, S., Gong, P., Wu, W., Davis, G., Qiu, D., Gu, X., Spear, R.: The use of remote sensing for predictive modeling of schistosomiasis in china. Photogram. Eng. Remote Sens. **68**(2), 167–174 (2002)
7. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases (1998)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning, vol. 1 (1993)
9. Pan, M., Wood, E.F.: Impact of accuracy, spatial availability, and revisit time of satellite-derived surface soil moisture in a multiscale ensemble data assimilation system. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **3**(1), 49–56 (2010)
10. Elshazly, H.I., Elkorany, A.M., Hassanien, A.E., Azar, A.T.: Ensemble classifiers for biomedical data: performance evaluation. In: 2013 8th International Conference on Computer Engineering & Systems (ICCES), Institute of Electrical & Electronics Engineers (IEEE), November 2013