A Cumulative Training Approach to Schistosomiasis Vector Density Prediction

A Cumulative Training Approach to Schistosomiasis Vector Density Prediction

Terence Fusco and Yaxin Bi

School of Computing and Mathematics
Ulster University
Newtownabbey, United Kingdom
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 epidemic disease vector and the rate of infection of the disease in any given community; it is this link that the project 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 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]. Whatever method is applied to the 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 C.T.A. 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 provided by our Chinese project partners 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.

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 to provide values for the environmental features present for each year. The number of common features from each year was 7 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 and distribution levels. When we can deduce which feature subsets are most influential on the density levels, we can make assertions on future SD classification based on these features and therefore provide important information to those concerned in order 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

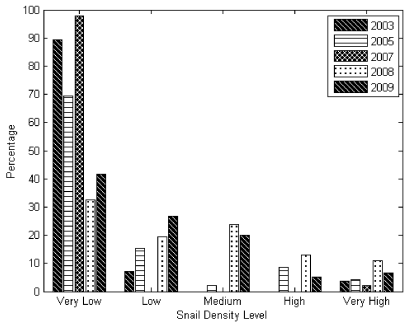


Fig. 1: Categorization of SD over Time

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.

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. The calculation is as follows:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad (1)$$

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 2 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 2 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 which makes 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 equally well with similar performance to one another 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 purposes. The SD data was separated at the beginning into 5 categories and the results are recorded in figure 1 and Table 1. 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.

Table 1: Average SD Values

Year of Collection	AVG SD	Normalised	Category
2003	0.72692	0.106948	Very Low
2005	0.878517	0.129335	Very low
2007	2.633028	0.388429	low
2008	1.396343	0.205804	low
2009	1.056396	0.155603	Very Low
Collective	1.01446	0.14941	Very Low

Table 2: Regression Results for Training and Testing Data

	2003	2005	2007	2008	2009
Linear Regression R^2 Value	0.325	0.590	0.734	0.808	0.699
SVR R^2 Value	0.052	0.221	0.732	0.691	0.506

2.1 Methods

In earth observation research, weather conditions directly affect the quality of satellite imagery observed which causes some values of environmental variables 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 can be due to the weather conditions that were present at the time of acquiring data from the satellites and 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 that would be 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 incomplete datasets 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. In the calculations below V represents the feature value of an instance.

WEKA replace missing value filter Replacing missing values with the mean and modal values from the remaining set for data Imputation.

Single PreSuccession Method Uses the previous and following values to replace the missing value.

$$v_i = \frac{v_{i-1} + v_{i+1}}{2} \tag{2}$$

Mean Single PreSuccession Method Uses the previous and following values to replace the missing value together with the average of the entire set.

$$v_i = \frac{v_{i-1} + v_{i+1} + \hat{v}}{3} \tag{3}$$

Double PreSuccession Method Uses the two previous and following values to replace the missing value.

$$v_i = \frac{v_{i-2} + v_{i-1} + v_{i+1} + v_{i+2}}{4} \tag{4}$$

Mean Double PreSuccession Method Uses the two previous and following values to replace the missing value together with the average of the entire set.

$$v_i = \frac{v_{i-2} + v_{i-1} + v_{i+1} + v_{i+2} + \hat{v}}{5} \tag{5}$$

We can see from Table 3 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.

2.2 Feature Assessment

– We want to evaluate the dataset to discover the relevance of each attribute to SD levels individually and as subsets of features.

Table 3: Data Imputation from 2007

Original Value	Weka Value Replacement	PreSuccession Method	Mean Single PreSuccession Method	Double PreSuccession Method	Mean Double PreSuccession Method
0.0348207	0.228452	0.25104015	0.168525222	0.012087575	0.060577163
0.128491	0.201003	0.10022725	0.253911833	0.123652625	0.068831075
-0.0844851	-0.083652	0.2363135	0.052187602	0.4010265	0.064255161
0.0244072	0.201003	0.09676155	0.105754767	0.02732925	0.0495664
-0.660181	-0.471425	0.0263615	-0.156374858	0.0796375	-0.083169715
0.242494	0.228452	0.3861414	0.213558972	0.1034562	0.071598343
-0.520896	-0.621769	0.386855	-0.090507671	0.3464385	-0.062387903
0.409524	0.6566	-0.006357	0.231914471	-0.141379	0.112144283
0.399665	0.344381	0.024851	0.122707944	-0.0248485	0.063684866
0.35545	0.6566	0.0067815	0.236293971	-0.119405	0.116539083
Average % Difference	145.13%	333.36%	226.81%	132.04%	32.58%

- To assess and rank the features of the data in each year 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 issue involved with remotely sensed 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.

2.3 Entropy

Entropy was carried out on each year of the data samples provided and results calculated in bits were recorded in the table to show the attributes that give the most information given the SD and distribution value. We can see from table 5 that the attributes TC_B and NDWI perform consistently across each year given the SD class. It can be seen that in years 07 and 08 that most of the attributes dropped in terms of their value to the class which is indicative of results of the overall SD classification accuracy attained in these years when tested. The entropy calculation is shown in equation 6 where entropy (H) and class is (C) [7].

$$H(C) \quad = \quad -\sum_{c \in C} p(c) \log_2 p(c) \tag{6}$$

2.4 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.

Table 4: Environmental attributes Entropy calculations

2003 Discretised	2005 Discretised	2007 Discretised	2008 Discretised	2009 Discretised
TC_B: 3.148 bits.	TC_B: 3.1666bits.	TC_B: 2.306bits.	TC_B: 2.426 bits.	TC_B: 3.103 bits.
TC_G: 2.962 bits.	TC_G: 3.234 bits.	TC_G: 2.897 bits.	TC_G: 2.852 bits.	TC_G: 3.2456 bits.
TC_W: 2.643 bits.	TC_W: 3.131 bits.	TC_W: 2.952 bits.	TC_W: 2.785 bits.	TC_W: 2.921 bits.
NDMI: 3.098 bits.	NDMI: 3.131bits.	NDMI: 1.020 bits.	NDMI: 2.988 bits.	NDMI: 3.095 bits.
NDVI: 2.693 bits.	NDVI: 3.218 bits.	NDVI: 2.162 bits.	NDVI: 2.104 bits.	NDVI: 2.752 bits.
MNDWI: 2.911 bits.	MNDWI: 3.045 bits.	MNDWI: 2.389bits.	MNDWI: 1.739 bits.	MNDWI: 2.855 bits.
NDWI: 3.156 bits.	NDWI: 3.081 bits.	NDWI: 0.156 bits.	NDWI: 2.058 bits.	NDWI: 2.939 bits.

Information Gain is a feature ranking approach that uses entropy to identify which feature in the dataset gains the most information. While entropy looks at the value of the attributes in relation to the set, Information Gain uses these results to assess the attributes when given the class information as a whole. This information is beneficial when carrying out analysis of features in a set to extract the most influential features in relation to their corresponding SD values. This can be of significant value in order to identify any corollary inferences with regards environmental features to SD levels. Once it can be identified which attributes have the most significant influence on the SD value, then it can be established for future experiments and predictions that these specific attributes are closely connected to high levels of SD.

The results show relative consistency with each year having similar positions for each of the attributes. The results are shown in Table 5. 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.

The information Gain calculation used is shown in equation 7 where entropy (H) is given of the class (C) given the attribute (A) [8]. Entropy and information gain are intrinsically linked as the decrease in the entropy of the class is a direct 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 [9].

$$H(C|A) \quad = \quad -\sum_{a \in A} p(a) - \sum_{c \in C} p(c|a) \log_2 p(c|a) \tag{7}$$

Table 5: 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

2.5 Correlation Analysis between Attributes

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 figure 2b and they indicate that data from 2008 is not in correlation with the alternate years as the trend line show us that the combination of the SD and the environmental attributes (X, Y) does not show correlation when compared 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.

$$P(x,y) \quad = \quad \frac{cov \quad (x,y)}{\sigma x \sigma y} \tag{8}$$

3 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 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 8/9 attributes given to experiment with. The most beneficial approach to dealing with

Table 6: 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

this issue is to use those attributes which are common to every year in order to make an equal 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 variant subsets of the data being used to discover temporal relationships within the dataset. This method of training will be referred to throughout this paper as the Cumulative Training Approach or CTA. The CTA method was used tested with 5 single classification methods to assess the best performance accuracy. We can see from figure 2a that year 2003 training data together 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.

We then 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 [10]. 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 7:

Table 7: CTA Ensemble Results

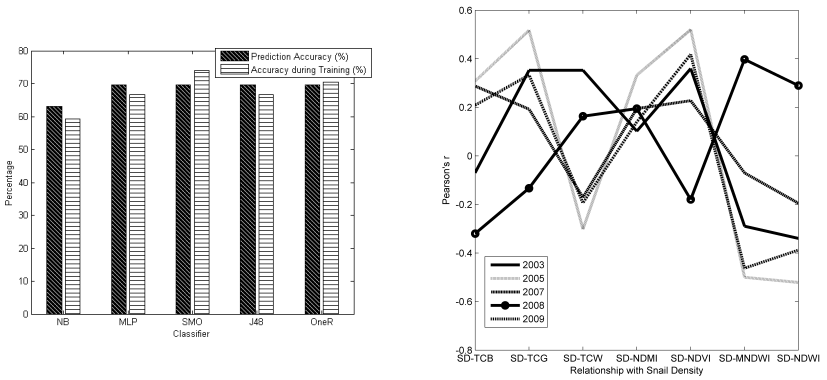
Testing Data	Training Data	Boosting		Bagging		Stacking	
		Training	Prediction	Training	Prediction	Training	Prediction
2003/05	2009	0.73973	0.48333	0.75342	0.5	0.71233	0.46667
2003/2005/2007	2009	0.52991	0.46667	0.50427	0.46667	0.55556	0.46667
2003/05/07/08	2009	0.55828	0.48333	0.5092	0.53333	0.4908	0.46667
2003/05	2008	0.73973	0.47826	0.75342	0.34783	0.71233	0.32609
2003/2005/2007	2008	0.52991	0.32609	0.50427	0.34783	0.55556	0.32609
2003/05	2007	0.73973	0.29545	0.75342	0.31818	0.71233	0.29545

4 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 remotely sensed data, which will be accessed in future research and experimentation. By testing effectiveness of the methods of replacement now, we can identify confidence in future replacement of data.

All experiments and collective research to date have become part of the cumulative training approach for Schistosomiasis vector density and distribution



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

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 a more significant 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 [11].

5 Future Work

Application for additional larger unlabelled datasets have been requested from our research partners and it is this data we hope to use together with our labelled data to provide semi-supervised cluster analysis. By using our labelled data as seeds to mix with new unlabelled data [12], we can separate the entire dataset into classes of low, medium and high SD based on the unlabelled data similarity and proximity to our labelled data. This will save a significant amount of time labelling the larger dataset and with the new labelled set, we can now implement machine learning classification methods and ensemble learning methods to provide accurate SD classification predictions based on environment factors.

Further to this testing process and analysis of relevant data, we propose to use unsupervised and semi-supervised k-means clustering to develop a deeper understanding of clustering patterns with the various feature combinations in this dataset. We will consider experimenting further using ensemble learning methods as these have shown to provide improved performance on individual classification methods [13].

These aspects and the methods detailed in this paper collectively contribute to the overarching theme and future application of a larger scale Cumulative Training Approach. This will ultimately will be implemented to achieve the greatest training potential possible for classification of SD for future testing and early warning of epidemic disease outbreak.

6 Acknowledgements

This work is partially supported by the Dragon 3 programme, a co-operation between the European Space Agency and the Ministry of Science and Technology of China. The authors would also like to acknowledge the Chinese partners at the Academy of Opto-Electronics, Chinese Academy of Sciences for making this data available for our research.

References

1. Ross, A.G.P., Sleigh, A.C., Li, Y., Davis, G.M., Williams, G.M., Jiang, Z., Feng, Z., Manus, D.P.M.C.: Schistosomiasis in the People s Republic of China : Prospects and Challenges for the 21st Century. **14**(2) (2001) 270–295
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), IEEE (oct 2014) 718–722
4. Sun, Q., Zhang, J., Zhou, J., Wu, L., Shan, Q.: Effect of poplar forest on snail control in dongting lake area. 3rd Intl. Conf. on Bioinformatics and Biomedical Engineering, iCBBE 2009 (2009) 1–5
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 (1) (jan) 216
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. **68**(2) (2002) 167–174
7. Almuallim, H., Dietterich, T.: Learning with Many Irrelevant Features. Proceedings of the ninth Nat. Conf. on Artificial intelligence **2**(Quinlan) (1991) 547–552
8. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases (1998)
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Volume 1. (1993)
10. 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing **3**(1) (mar 2010) 49–56
11. 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). (2013) 184–189
12. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised Clustering by Seeding. ICML 2002 (July) (2002) 19–26
13. Jurek, A., Nugent, C.: A Cluster-Based Classifier Ensemble as an Alternative to the Nearest Neighbor Ensemble. In: 2012 IEEE 24th Intl. Conf. on Tools with Artificial Intelligence. Volume 1., IEEE (nov 2012) 1100–1105