International Journal of Machine Learning and Cybernetics (2020) 11:1159–1178
https://doi.org/10.1007/s13042-019-01029-x

ORIGINAL ARTICLE

Data mining and machine learning approaches for prediction modelling of *schistosomiasis* disease vectors

Epidemic disease prediction modelling

Terence Fusco[1] · Yaxin Bi[1] · Haiying Wang[1] · Fiona Browne[1]

Received: 27 August 2018 / Accepted: 29 October 2019 / Published online: 18 November 2019

Abstract
This research presents viable solutions for prediction modelling of *schistosomiasis* disease based on vector density. Novel training models proposed in this work aim to address various aspects of interest in the artificial intelligence applications domain. Topics discussed include data imputation, semi-supervised labelling and synthetic instance simulation when using sparse training data. Innovative semi-supervised ensemble learning paradigms are proposed focusing on labelling threshold selection and stringency of classification confidence levels. A regression-correlation combination (RCC) data imputation method is also introduced for handling of partially complete training data. Results presented in this work show data imputation precision improvement over benchmark value replacement using proposed RCC on 70% of test cases. Proposed novel incremental transductive models such as ITSVM have provided interesting findings based on threshold constraints outperforming standard SVM application on 21% of test cases and can be applied with alternative environment-based epidemic disease domains. The proposed incremental transductive ensemble approach model enables the combination of complimentary algorithms to provide labelling for unlabelled vector density instances. Liberal (LTA) and strict training approaches provided varied results with LTA outperforming Stacking ensemble on 29.1% of test cases. Proposed novel synthetic minority over-sampling technique (SMOTE) equilibrium approach has yielded subtle classification performance increases which can be further interrogated to assess classification performance and efficiency relationships with synthetic instance generation.

Keywords Disease prediction modelling · Data imputation · Synthetic data simulation · Schistosomiasis · SMOTE · Incremental transductive approaches

1 Introduction

Epidemic diseases are becoming more prevalent in developing countries and serious consideration is being given to purposeful ways of preventing them by utilising satellite technology combined with ecological information [1]. Artificial intelligence applications for addressing disease preparation and control have yielded promising results in recent years. Promising techniques in the bioinformatics domain include using remote sensing (RS) and earth observation (EO) techniques to provide ways in which to investigate land characteristics remotely [2]. These RS and EO methods help to reach some of the most rural localities and provide early risk warning to those in remote areas that would have been extremely difficult previously. EO methods enable the construction of geographical information systems (GIS) that can provide detailed mapping of at-risk areas. These mapped areas can be dissected more efficiently to extract the required environment information for assessment [3]. Epidemic diseases are defined as those which spread rapidly and at the same time therefore it is crucial to investigate potential indicators of transmission likelihood in a given area for providing advanced warning information to those at risk [4].

EO is a domain which has attracted attention in land and environment monitoring as it utilizes satellite technology to analyse occurrences on the earth's surface [5]. The ability to remotely sense land characteristics of a given area, enables

✉ Terence Fusco
fusco-t@ulster.ac.uk

[1] Faculty of Computing and Engineering, University of Ulster, Newtownabbey, UK

research to be conducted on a global scale which can have a considerable effect on problem solving for disease prevention. There has been much improvement in EO technology in recent years resulting in higher definition images from which to extract environment features with more accuracy. Disease prediction models are being applied in many current research studies to detect early warning information for disease outbreak due to success of machine learning applications with satellite image extraction techniques [6]. Using RS and EO approaches provides an insight into the necessary land and hydro-logical information that indicates likelihood of high disease vector density which highlights higher epidemic disease transmission risk [7]. In terms of specific disease in the context of this research, our focus is on vector-borne diseases such as *schistosomiasis*, *dengue fever* and *malaria*. These are the three most prevalent and dangerous epidemic diseases in terms of numbers of infection as stated by the World Health Organization.

In this paper, the focus is on *schistosomiasis* epidemic disease (SED) and in particular the host vector freshwater snail [8]. The problem being addressed concerns improving classification accuracy of SD levels and increasing training potential of a sparse real-world sample. The methodology presented refers to proposed approaches for accurate prediction modelling of disease vector density indicating likelihood of transmission risk in a selected area. This research focuses on the classification problem of predicting the density levels of schistosomiases disease vectors in order to identify areas of high disease risk. The input variables in this work are the environment factors used for predicting snail density such as soil moisture, vegetation and tasselled cap indices. The output variable refers to the snail density level classification i.e. 'Low', 'Medium' and 'High'. Consequently, if the environment conditions of a particular study area are similar to those with proven high snail density then we can infer that the surrounding freshwater area is of potential risk and provide communities with a health warning. Our approach is to assess the risk of disease transmission using environment-based data as training seed when making predictions for early warning detection of *schistosomiasis* transmission. Vector density and distribution levels indicate an increased or reduced likelihood of potential infection from freshwater sources which are assessed using labelled environment information from relevant areas [9]. There are many lakes of interest when analysing freshwater snail distribution for SED research in China. These include areas in proximity to Dongting Lake and neighbouring Poyang Lake which can be used to assess freshwater snail distribution in different regions of China [10]. Dynamic land conditions present at any one time on earth can tell a lot about what can happen in terms of natural disasters such as earthquakes and they can also indicate conditions which are most conducive for disease vectors to thrive [11].

Composite environment feature combinations can have significant influence on propagation of the freshwater snail host of SED. For example, it is the case that freshwater snails prefer warmth and moisture in their habitat for optimum breeding conditions. This indicates that if soil moisture levels and climate conditions are favourable in a relevant area then there is increased chance of high snail distribution and density in that area [12]. The only available treatment for those infected with SED is the drug Praziquantel which can be taken once infected and therefore cannot be used for prevention [13]. In most high-risk areas of epidemic disease infection, similar climate conditions are present as well as socio-economic circumstances. Studies show that due to the preferred temperatures for disease vectors to flourish, the majority of epidemic diseases cases investigated can be found in parts of Asia, Africa and South America which provide ideal surroundings for vectors to multiply [14].

One of the problems being addressed with this work concerns the lack of real-world data and difficulty in attaining training samples for use with developing disease prediction models. In addition, problems exist with partially complete data for use with machine learning methods due to lack of satellite image clarity. These problem areas have yielded perennial issues with satellite image extraction techniques and disease prediction modelling accuracy. There exists somewhat of a dichotomy in the realm of data mining and machine learning domains as to how best to approach the classification process in terms of training data volume and validity. The paradoxical research options of using a sparse set of real-world data being expensive and time-consuming to collect, compared with using satellite image extraction of environment features yielding vast repositories of training data, has proven to be a difficult issue to clarify. It is the assertion of this research that the most appropriate approach is to apply machine learning methods using existing sparse field-survey data for building prediction models. Field-survey data is combined with satellite extracted environment images of the same area for corroborating classification experiments. To improve scope of the machine learning process in future, the aim is to use these studies for SD classification of future RS generated unlabelled data in order to make accurate predictions of disease transmission likelihood in a specified area. Solutions proposed in this paper aim to address and improve on issues highlighted while taking into consideration previous work in disease prediction modelling and data imputation [15, 16].

Firstly, a method was developed for value replacement namely, the regression-correlation combination (RCC). This model was built on foundations of initial research involving a cumulative training solution when using a sparse training sample [17]. Incomplete training data reduces machine learning potential hence the requirement for viable data imputation methods to handle missing values with precision

and consistency. Providing a credible imputation model for training purposes will improve reliability and confidence of experiment results without distorting the original set.

Then further development was made to the previously proposed incremental transductive methods [18] which include single and multi-classifier approaches for labelling snail density (SD) classes using an incremental, semi-supervised learning process. The rationale for this work was to analyse the accuracy and efficiency of labelling potential when using sparse data samples. Data labelling in this context is used to assess effectiveness of environment values to accurately classify SD category of a specified area consequently highlighting likelihood potential of *schistosomiasis* transmission in that area.

Finally, the focus was on data simulation as a way of assessing exponential increases in training instance volume and whether classification results achieved were markedly different from original instance numbers for each year. Training data was first processed using synthetic minority over-sampling technique (SMOTE) by applying equal SD class numbers to each test parameter before synthetic instance increases to assess larger sample performance improvement and avoid potential over-fitting issues [19]. We propose that ensuring an equal number of SD instances before applying SMOTE with exponential increases; may be advantageous to classification results without providing skewed results. Imbalanced data is a problem faced due to the nature of this study area in which there can be a large distribution of disease vectors in a study area at either end of the scale. This may skew results and give a misrepresentation of the study area as a whole therefore, our aim is to build synthetic prediction models which provide a varied data pool for applying machine learning approaches when using a limited training sample. A flowchart depicting these applied methods is shown in Fig. 1 with details discussed further in this research along with benchmark results. The proposed methods in this research are developed to address this niche disease study area with results having implications on the wider epidemic prediction field for *schistosomiasis* disease. The novelty of methods proposed and used in each of the experiments encompass the synthesizing of incremental machine learning with transductive reasoning for the training sample to learn while actively labelling instances in the set. The raw training data supplied for this work with missing values required investigation into imputation approaches that would be cognisant of the original data sample for application to future partial training data.

The main issues being addressed with this research and technique contributions presented are as follows:

1. Handling of partially complete training data which can have a detrimental impact on the machine learning classification process especially when using a sparse training sample.A novel RCC method is proposed to compare performance with standard WEKA value replacement and previously proposed mean double pre-succession value (MDPV) replacement. Benefits of providing highly accurate replacement values when faced with incomplete training data will enable greater confidence in the efficacy of disease likelihood predictions and increase the training pool for improved training potential.
2. Inconsistent prediction accuracy performance for future snail density and distribution levels when provided with environment feature values. A novel approach namely, incremental transductive ensemble approach (ITEA) models are proposed to provide Snail Density class labels. If we can accurately classify labels from future environment feature values then informed and actionable *schistosomiasis* likelihood predictions for early warning and preparation can be provided.
3. Lack of real-world sample data available due to expense and labour requirements poses a problem for classification and prediction performance. A synthetic data simulation application is presented in which SMOTE is

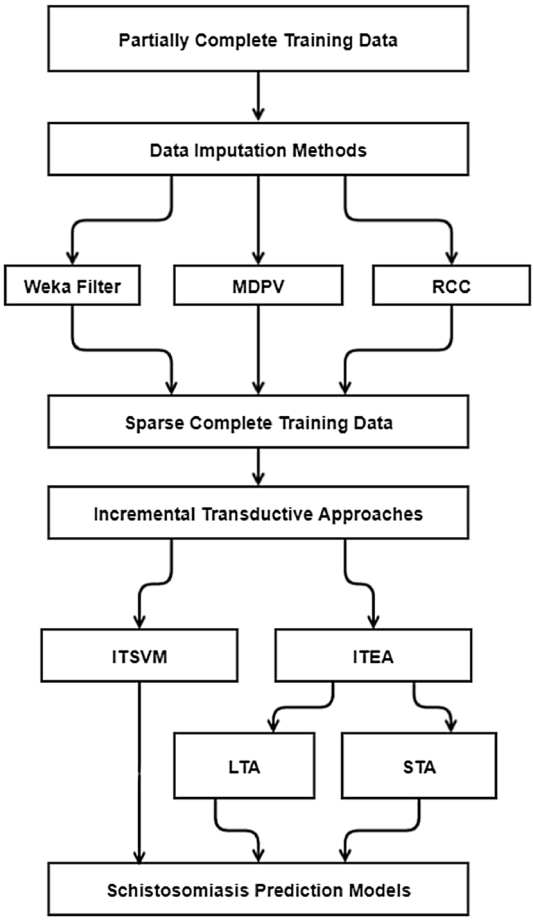


Fig. 1 Prediction model flowchart

applied after pre-processing raw sample data then equalising the number of classes to avoid potential imbalance and over-fitting issues. This provides empirical evidence as to the benefits of using sparse real-world data for training as opposed to a much larger pool of machine generated data based on the original training set. Implications of these results will inform future research as to whether unlabelled big data is preferable over sparse real-world data for classification and prediction modelling.

1.1 Related work

For disease classification and prediction in general, successful research methods have tended to apply revisions of traditional standalone algorithms for achieving highest vector classification accuracy [20]. Variations of existing classification methods including Naive Bayes (NB), are developed with the aim of improving on shortfalls of standard application of these classifiers making them more adaptable to particular environment-based or human genetic sample training data [21].

Relevant research has been conducted into severe acute respiratory disease (SARS) using spatial distribution clustering together with socio-economic factors and support vector machine (SVM) classifier [22]. Results show strong correlation between environmental indices and disease transmission rates when data was analysed from the outbreak in China of 2002/03. This particular study was carried out across 31 provinces and provides a balanced representation of variable climate and terrain parameters with classification based on disease cases in each area. Subtle modification of algorithms developed for multi-variate classification and labelling purposes can provide much improved results and a more successful approach to classification pertaining to disease infection [23]. Research has shown that some of the most successful machine learning techniques used in disease prediction modelling involve using suitable ensemble learning methods. This is due to the accumulation in relevant studies of improvement on classification accuracy performance over single classifier solutions [24].

Consideration has been given to many other aspects of research and analysis in disease related fields including current pre-processing methods which aim to boost performance of existing algorithms and improve efficiency by removal of outliers in the data [25]. A key focus of this research in particular is to synthesize and assimilate available work in this area with assessment of field research sample data used for experiment purposes. Evidence of related disease prevention work and current state of art suggests that customising traditional algorithms is the most promising approach to deal with the issue being addressed. Classification results vary as expected depending on parameters, sample size and nature of disease with *Arbovirus* classification reaching 90% accuracy during experiments [26]. If this research can be further developed by applying proposed approaches to a variety of epidemic disease prediction problems; then it may be validated as being applicable to multiple environment-based diseases and prove a useful resource for pre-emptive preparation for combating infection of these diseases globally.

2 Materials

When applying machine learning techniques, the assumption is that a larger training pool provides greater resources for making accurate classification decisions. This may not necessarily provide authentic optimal results in contrast to using a succinct real-world dataset acquired using manual field research methods. Due to technological advancements, it is possible to access huge caches of satellite imagery from around the globe from many spatio-temporal parameters for environment feature extraction purposes however, field survey samples are much more scarce and difficult to acquire. In this collaborative research study, satellite images are used by research partners for image recognition and feature extraction using RGB-scaled image recognition which are then processed with environment features and corresponding field-survey values provided for comparative analysis. The resulting raw sample data is then processed and applied to train machine learning algorithms for making SD classifications.

Classifiers selected in this research were chosen due to previous success and popularity in this research domain. The well-established classifiers used are NB, SVM, J48 decision tree and multi-layer perceptron (MLP) with the objective of providing variety in experiment as well as identifying strengths and weaknesses in relation to classification of each set. There were many other possible algorithms to apply for initial research however, those used in this study have proved to perform well, specifically in disease related research There are many epidemiology studies using pre-processing models currently available that have specific application for various disease classification approaches. They have shown to perform well in conjunction with existing traditional algorithms such as neural networks, decision trees, SVM and Bayesian probability [25]. Indeed, it is noticeable that SVM performs well in comparison with traditional algorithms when applied for arbovirus epidemic disease prediction modelling with 90% classification accuracy achieved [26]. This study however, was achieved using 5000 instances for training purposes whereas a much more limited data sample is available for performing machine learning experiments in this work. There are also interesting results using spatial clustering and geographical weighted regression models which take into consideration temperature readings, soil moisture and elevation in relation to SD levels which resulted in 68.93% snail dispersion classification [45].

A notable and novel approach is to use a fusion of methods to achieve the intended research goal. This approach has

the possibility of combining the optimal performances of a range of different algorithms which then work together to classify the target object. In one particular study, researchers were seeking to identify the contents of a particular web page by fusing textual and visual components together and then using the outputs with application of the NB algorithm to then classify whether or not there was phishing involved on the web page [44]. This approach is worth considering when studying on a multi-variate analysis such as this current study as the use of differing approaches to achieve the same goal through a process of synthesizing identifying aspects in the dataset. For instance, if there were specific methods that could predict disease vector density that could be combined with a method that could accurately predict the future climate or weather conditions in a study area then it would beneficial to fuse the method outputs together to identify potential disease risk in that area.

The overall aim is to provide highly accurate classification results for making credible future SD predictions. Sample training data used in all experiments was collected by Chinese project partners over the course of 6 years ranging from 2003 to 2009 in the Dongting Lake region of Hu'nan province of China[1]. Training data was collected over a number of years in the same season of year for spatio-temporal continuity and comparison. Labelled environment feature data was provided by research partners who are domain experts in disease prevention and control. The raw experiment data was then preprocessed and prepared for use with classification methods.

This research is a multi-disciplinary project with the European Space Agency (ESA) partners providing satellite information and Chinese partners at the Academy of Opto-electronics providing the raw environment training data extracted from satellite images. Our particular focus in the project relates to the assessment and analysis of all raw data resources with the aim of building viable prediction models. Climate and environment conditions are changeable year on year therefore, thorough investigatory techniques are required for assessing data samples from each year in order to interrogate relationships between SD levels and yearly fluctuating environment values. In each of the datasets, instances are represented by environmental factors extracted using RS satellite techniques together with field survey samples. The combination of environment image extraction and field-survey samples provide training data for making predictions of future snail density without need for further samples to be collected manually. Environment features are recorded using pixel colour values with each regional study area being an object which is representational of the surrounding area. Training samples were acquired on a year by year basis in the same area and season for continuity and empirical analysis.

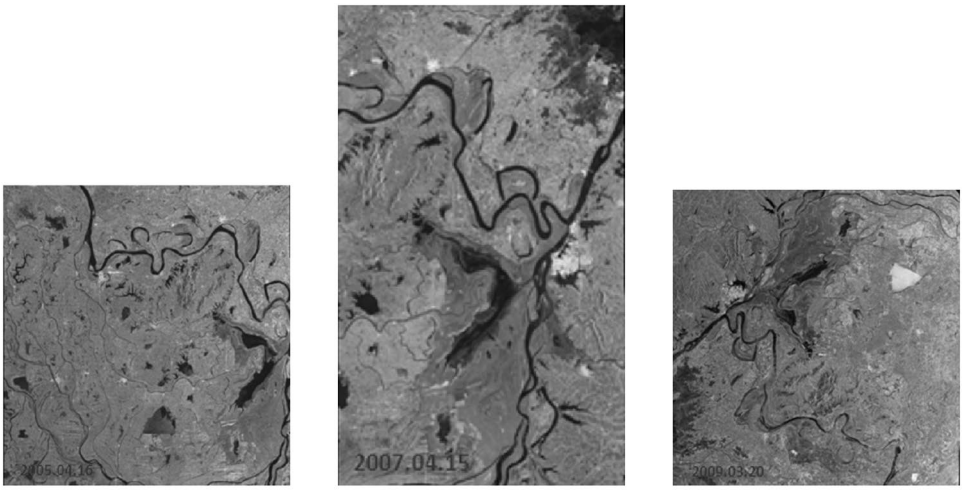
For the purposes of field survey research, the study area around Dongting Lake was divided into 0.11 (m^2) sections and labelled by experts with a serial number prior to using specialist apparatus for measuring environment feature values. Freshwater snails in each of the selected areas were carefully collected and counted before recording all information for further analysis. Examples of satellite imagery used in this work are shown in Fig. 2. They represent satellite viewpoints taken over the Yangtze River with Dongting Lake shown as a large body of water. These particular images were extracted during the months of March and April in years 2005, 07 and 09 with the 2009 image being taken from an alternative viewing perspective.

Data used in this work were originally imbalanced and varying in environment attribute numbers. To address this issue, environment attributes were aligned to provide a common attribute training sample for comparative analysis across each year as shown in Table 1. Once common attributes were aligned, each training set was normalised using a Softmax function which constricts snail density value range into a vector between 0 and 1. The values were then discretised in order to allocate them with a density category in preparation for classification experiments. During pre-processing of raw data, every consideration has been made to carefully handle the data in order to avoid distortion from the original survey sample.

The total number of common environment attributes in the training sample is seven with the collective instance number being 223 in addition to SD values. These are the constant training criteria used for each of the experiments in this research. Environment features included for each instance together with corresponding SD values are as follows: Tasselled Cap Brightness (TCB), Tasselled Cap Greenness TCG, Tasselled Cap Wetness (TCW), Modified Normalised Difference Water Index (MNDWI), Normalised Difference Moisture Index (NDMI), Normalised Difference Vegetation Index (NDVI) and Normalised Difference Water Index (NDWI). The limited field survey data samples acquired by research partners are as follows: 2003–27 instances, 2005–46 instances, 2007–44 instances, 2008–46 instances, 2009–60 instances. While the data set is relatively small in data mining terms, it provides a solid basis on which to form initial opinions and observations as to which features or combination of features have the strongest influence on SD levels. When deducing which feature subsets are most influential to the SD levels, assertions can be made on future SD classification and therefore provide the most actionable and important information to those concerned for preventative measures to be implemented. Each data set listed including the collective set were applied to proposed approaches

[1] This data used in this research was provided by the European Space Agency and partners at the Academy of Opto-Electronics, Chinese Academy of Sciences, China

Fig. 2 EO data sample



(a) 2005.04.16 **(b)** 2007.04.15 **(c)** 2009.03.20

in this work for comparative analysis and validation of methods. A sample snapshot of raw training data is shown in Table 1; values are rounded to two decimal places for this example, however actual recorded data used in experiments includes up to five decimal places. All aforementioned environment features are relevant to the freshwater snail desired habitat and have shown to be indicators of potential heightened disease risk [27]. Using relevant analytical approaches helps to ascertain if there exists any correlation between individual or component environment features and high levels of SD for prediction purposes.

3 Methodology

The a-priory assumption of this research is that snail density and distribution levels are derivative of and directly linked to climate and environment conditions in a given area. Initial inference is that high levels of vegetation and moisture present in a particular area can greatly increase breeding potential of the freshwater snail and provide a likely indication of snail presence in particular areas.

3.1 Proposed approach

This research is addressed from a vector class labelling standpoint in which the aim is to provide SD labels for training instances using field survey data. The real-world collected data in this research will is used as training seed for labelling future RS generated data which is then used for making predictions. A transductive labelling approach was initially applied [28] with an incremental element added which meant that any new instances capable of being labelled could then be added to the original data pool with the aim of increasing the chance of providing labels in successive increment phases of the labelling

Table 1 Common environment attributes

Attribute	Name	Description	Reason For use
TC_B	Tasselled Cap Brightness	The Brightness value of a pixel in an image	(Band 1) Measure of soil
TC_G	Tasselled Cap Greenness	The Greenness value of a pixel in an image	(Band 2) Measure of vegetation
TC_W	Tasselled Cap Wetness	The Wetness value of a pixel in an image	(Band 3) Interrelationship of soil and canopy moisture
MNDWI	Modified NDWI	Modified NDWI uses (MIR) middle infra-red instead of (NIR) near infra-red remote sensing	Modified NDWI can enhance open water features while reducing land/vegetation/soil noise
NDMI	Normalised-Difference Moisture Index	Moisture Index	Used to assess whether the target being observed contains much soil moisture
NDVI	Normalised-Difference Vegetation Index	Vegetation Index	Used to analyse whether the target being observed contains live green vegetation
NDWI	Normalised-Difference Water Index	Water Index	Water index that uses near infra-red band. The image is not modified and can include many other factors which can confuse the reading

process. Proposed LTA and STA are an extension of the incremental transductive process which apply multi-classifier options for labelling. These ensemble approaches were a natural progression when using a multi-classifier solution for labelling as they provide options to determine the required confidence level as well as distinguishing agreement between classifiers before a class label is provided. The ITEA methods are a development of previous work [18] which detailed initial ideas and rationale to introduce the multi-classifier option into the incremental process in order to provide a choice for labelling when using multiple algorithms. The algorithm described in the Pseudocode essentially shows the procedures for LTA and STA proposed approaches which are similar in their structure. The LTA method declares that for unlabelled instances, if the probability confidence threshold is met by any of the classifiers; then that corresponding SD label is provided to the instance. The STA method is similar however, it requires all of the classifiers otherwise no label is provided to the instance. The selection of the most appropriate ensemble approach depends on the stringent requirements of a particular research study.

Pseudocode for the LTA and STA algorithms used in the ITEA can be seen in Algorithm 1:

- Where L is the set of labelled instances
- U is the set of unlabelled instances,
- l is a labelled instance
- u is an unlabelled instance
- n as the number of iterations with no labelled instance
- γ is the confidence threshold for classifiers
- ϕ is the number of passive iterations permitted without any labelling taking place prior to ending the process

While methods in this paper are specific to SED, they may also be compatible for alternative environment-based epidemic disease prediction modelling that relies on limited sample data to achieve intended results. It is evident for research problems of this nature that spatial and temporal parameters can prove difficult to access on demand for research corroboration. This consideration as well as regular fluctuations in climactic conditions make like for like analysis of little use in some cases).

Algorithm 1 Incremental Transductive Ensemble Approaches

procedure LTA$(L, U, l, u, n, \gamma, \phi)$
  while $U \neq \emptyset \vee n \geq \phi$ **do**
    $n = 0$
    $TrainClassifiers$
    for $u \in U$ **do**
      if $P_1 \geq \gamma \vee P_2 \geq \gamma \vee P_3 \geq \gamma$ **then**
        $L \leftarrow u$
        $n = 0$
      else
        $n + 1$
      end if
    end for
  end while
end procedure

procedure STA$(L, U, l, u, n, \gamma, \phi)$
  while $U \neq \emptyset \vee n \geq \phi$ **do**
    $n = 0$
    $TrainClassifiers$
    for $u \in U$ **do**
      if $P_1 \geq \gamma \wedge P_2 \geq \gamma \wedge P_3 \geq \gamma$ **then**
        $L \leftarrow u$
        $n = 0$
      else
        $n + 1$
      end if
    end for
  end while
end procedure

3.2 Classification algorithms

Four core classifiers were applied to assess the predictive performance for labelling of SD classes including NB, SVM, J48 Decision Tree (J48) and MLP. These were selected due to the positive results received when applied to the disease prediction problem domain and also due to the variety in each approach which yields greater balance for experiment conditions. The classification is focused on three SD classes of 'Low', 'Medium' and 'High' making it a multi-variate analysis problem that is used to provide indication of disease likelihood rather than binary dense or not for improving disease likelihood analysis. If these algorithms perform well during testing then they can be considered for modifying to increase classification effectiveness with this research. Classifiers used in each of the experiments in these studies to assess the distribution and density of *schistosomiasis* vector levels. NB is a probabilistic algorithm which aims to classify data instances without bias based on the vector class properties. The NB classifier was found to provide consistent performance across the SED prediction domain however, it struggles to achieve the highest classification accuracy results when compared with other algorithms. SVM classification splits the data using a hyper plane which then deduces the class and instance it should reside in. SVM has shown to provide increased classification accuracy over NB in disease prediction research however NB is relatively more consistent across varied datasets [26]. Modified versions of SVM have been widely used with success in the area of disease predictions in epidemiology studies [29] thus it was deemed suitable and applied it with the previous method. Based on the Java implementation of the C4.5 algorithm, J48 creates rules for classification based on the information present in the set [30]. During experiments it was shown to perform well when compared with the other selected algorithms in many instances with higher classification accuracy percentages. MLP is a classification tool which functions as a feed-forward neural network and has shown promising results in data classification [31]. One of the advantages of its application is that it continually processes data in order to distinguish classes. In this study, ten layers are used with five neurons hidden and three output neurons. In addition to aforementioned classifiers, a LibLinear algorithm was applied due to positive performance in similar multiclass problem research [32] and a novel feature ranking approach namely max-relevance-max-distance for comparison with our proposed methods [33].

In Eq. (1) requirements for assigning SD class labels to each instance is detailed. It is the case that during the labelling process, a label is assigned to a class if and only if the probability of the class of the instance is greater than or equal to the probability of the class c given the instance i.

$$L = c_i \quad iff \quad p(c_i) \geq p(c|i) \tag{1}$$

The approaches used in this paper were applied to label a selection of data that was collected using field survey research and remotely sensed image extraction of environment features. Experiments have been carried out on each year of training data from 2003 to 2009, which was provided by research project partners at the ESA and Academy of Opto-Electronics in China as detailed in the "Materials" section. SD classification accuracy for each year has been analysed to assess proposed methods as well as the *f-measure* which uses precision and recall for recording results and provides a full picture of classifier performance than accuracy results alone.

3.3 Data imputation methods

When analysing training data provided by partners at the ESA and the Academy of Opto-electronics in China, some samples were only partially complete. This is a common problem when analysing satellite images and is related to the climate conditions at the time of image acquisition [34]. For example, rain or heavy fog will reduce visibility from space rendering poor resolution from which to extract environment information from. This research issue is important to address for these occasions as a partially complete dataset can considerably reduce machine learning effectiveness for classification and making predictions. The necessity for addressing missing data instances is due to the nature of classification for these disease vectors and difficulty in attaining primary field survey data. Larger training sets are desirable for use in prediction modelling due to increased training and classification potential so the aim is to retain as much collected sample data as possible especially in the case of learning with limited resources. Simply removing instances which are incomplete would likely result in higher SD classification accuracy in many cases however it would also distort the sample data and misrepresent the selected study area. This would not be indicative of the state of environment therefore the objective for real-world samples is to interrogate authentic environment data as much as possible to extract the prescient information. Common imputation methods for dealing with this problem tend to use mean attribute values from the original set, however the aim of this work is to provide a more bespoke dataset inclusive solution that outperforms standard methods when replacing actual values:

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

Data imputation experiments were conducted empirically by manually removing a random number of instance attributes from each year then replacing them using a combination of methods which consisted of applying:

- WEKA missing value replacement filter (Version 3.7).
- The previously proposed MDPV.
- A novel RCC which uses correlation across the existing data then regression to replace the missing values.
- Iterative Robust model-based Imputation (IRMI) which uses sequential and iterative approaches for imputing missing data from a variety of data types.

The RCC method was proposed to incorporate some of the CTA framework into a unified tool for data imputation in partially complete data. It performs regression analysis on all pairs of attributes and uses the pair with the highest R^2 value to impute the missing value. The MDPV method proposed previously is shown in Eq. (2) with V_i representing the imputed value, V_{i-2} representing the neighbouring value two places to the left of the imputed value, V_{i-1} is the neighbouring value one space to the left with V_{i+1} and V_{i+2} representing the corresponding values to the right of the imputed value which are then all divided by five to provide the replacement value. IRMI was introduced as a way of evaluating the performance of the proposed RCC approach with a contemporary method used with success in pre-processing research [46]. Using a sequential and iterative approach can be of interest with limited training data to help inform the imputation process and with this IRMI method, a variable is used for each iteration as the response variable while the other variables are used regressively [47].

3.4 Incremental transductive approaches

Much research has been carried out using various labelling mechanisms including supervised, semi-supervised and unsupervised approaches. These methods have shown to be successful in providing accurate class labels to a variety of different machine learning problems. Limited availability of real-world training samples is a common problem in this research discipline and one that is being continually addressed to find a satisfactory solution [35]. Of course, the learning approach that applied is dependable upon the research problem being addressing. Based on the requirement to make predictions using limited training samples as seed for future classification, this indicates using a combination of labelled and unlabelled data with this research which is deemed most suited to semi-supervised machine learning [36]. It is the assertion of this research that semi-supervised approaches are most suitable to apply to this problem domain due to their ability to provide sufficient solutions when considering classification problems with limited training samples [37]. A key contribution of this work is to make predictions on future epidemic disease likelihood based on the environment factors present in a particular area. Taking into consideration acquired field-survey samples over a number of years for making future unlabelled predictions, semi-supervised learning is the most appropriate paradigm to pursue. In addition to this, current research using semi-supervised learning and classification is progressing and providing consistently promising results. Introducing an incremental training element to a semi-supervised labelling process provides a more proactive approach in terms of the machine learning applied for accurate SD classification. This active learning process has the capacity to inform and update the machine process while simultaneously providing class labels [38]. The rationale is to maximize the machine learning capacity before labelling SD instances and has proved to be viable and effective strategy for semi-supervised research application [39].

ITEA performance in this paper is evaluated with comparative analysis using the same selection of algorithms and standard Stacking ensemble classification. ITEA was proposed and implemented to inform the research process for future experiments and demonstrate how selected classifiers behave under different constraints when labelling SD classes. Application of ITEA methods used in addressing vector classification or with alternative related research issues is dependable on user requirements and whether the necessity for providing a class label is prioritised over a more stringent labelling process; for example a scenario where an efficient approximation or probability is required rather than a critical piece of information. Experiments were conducted using well-established algorithms in the bioinformatics research domain including NB, SVM, J48 and multilayer perceptron with application of WEKA machine learning software. Classifiers were combined using the default stacking meta-learner implemented in WEKA and compared with proposed liberal (LTA) and strict (STA) training approaches. As part of the incremental transductive models proposed, a set of constraints must be selected as well the confidence threshold to which any label can be assigned by the classifiers. Then the passive iteration threshold must be set that stops the labelling process when the active learning process can no longer provide new labels for SD classes. The confidence level was set at 85% for each of the tests and the same four algorithms were used for each training set.

3.5 Incremental transductive support vector machine

Previous research has shown modified versions of SVM application to provide promising results in the area of epidemic disease prediction modelling. In light of this, it was decided to implement SVM as part of a semi-supervised incremental approach with the aim of developing an active learning process that can provide SD class labels with high degrees of confidence and consistency. Incremental transductive SVM (ITSVM) was previously proposed as a semi-supervised method for labelling SD classes and applies the

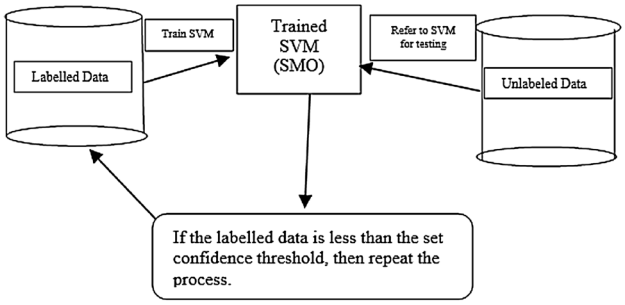


Fig. 3 ITSVM process

SVM classifier incrementally with newly labelled instances being added to the original sample for increasing training potential. Benchmark results in these initial studies comparing ITSVM with standard SVM application resulted in SD labelling improvement on 21% of test cases [17]. It is apparent from the ITSVM process in Fig. 3 that existing instances are used with the SVM classifier for assigning SD class labels in a semi-supervised process. Newly labelled instances are then added to the training data thus increasing labelling potential in an incremental active learning process.

3.6 ITEA

In addition to ITSVM application, the labelling process was further developed by applying a multi-classifier labelling approach. The basis for this is that ensemble classifier solutions have shown to provide increased classification performance accuracy than standalone algorithms [24]. This being the case, novel ITEA approaches were developed to provide SD class labels using a combination of multi-classifier learning methods. As part of the ITEA, two implementations are derived which are LTA and STA training approaches. These approaches are designed to provide bespoke labelling processes using multi-classifiers solutions for the providence of SD class labelling when required to a specified confidence level. The rationale for these ITEA methods was to provide a classification process with the caveat of label confidence selection and stringency of corroboration between chosen classifiers. Applying these methods for problems in a case specific and selective manner may provide the most appropriate approach for that particular sparse data paradigm. Vector labelling methods in this research are not limited to *schistosomiasis* disease alone but can be applied for alternative environment-based disease prediction models that use machine learning solutions. The first ITEA sub-method is a LTA which has a labelling requirement of any single selected classifier from those chosen which is capable of assigning a class label to the selected confidence level. Using the LTA method requires selecting a number of classifiers to build an SD labelling model with the basic constraint of a one classifier assignment strategy. Prior to running the labelling process, a pre-defined confidence level is decided then during the incremental process the liberal aspect of the method requires a single classifier to meet the desired confidence level in order for a class label to be assigned. The STA is similar process to the LTA with the confidence selection and classifiers pre-defined but the point at which it differs concerns label assigning constraints. With the STA, all classifiers selected must meet the required confidence levels and assign a label to the selected class in agreement otherwise the instance remains unlabelled.

3.7 Synthetic data simulation

This section is focused on data simulation and gives an insight into the validity of using a snapshot sample of environment data as opposed to construction of an increased synthetic dataset for epidemic disease vector classification. Synthetic data instances used in experiments are created based on original real-world training data provided by our research partners. Raw training sample data was pre-processed before filtering the data then applying SMOTE. The rationale is to first increase the sample size to attain an equal number of classes in each set before incrementally increasing synthetic instance numbers to the stipulated size.

The SMOTE technique generates an increased number of synthetic data instances based on the original dataset provided. SMOTE constructs synthetic instances of data which aim to improve classifier performance by providing a balance of over-sampling the minority class and under-sampling the majority in a way that seeks to reduce the loss ratio during classification [40]. SMOTE is a method traditionally used for over-sampling and data imbalance issues. The purpose of SMOTE application with this research is to address the issue of sparsity of data for training purposes. The aim is to assess the effectiveness of increasing instances of data on classification performance in contrast to real-world limited samples. As part of this research, SD class numbers were equalised in order to avoid skewing results due to the fact that there is varied distribution in the target study area which potentially provides imbalance and over-fitting. The SMOTE Equilibrium technique was implemented by pre-processing raw field survey data then equalising the number of SD classes from each year. This synthetic instance generation should result in a larger training set with less over-fitting for applying selected classification methods.

The aim is to utilise this approach with the training data for comparative analysis purposes during the classification and prediction modelling process [41]. Ultimately, if applying SMOTE can greatly improve classification performance over original data performance, then the process can be optimised to provide a bespoke model for use with this research issue. The rationale behind proposing SMOTE is

based on the fact that although there is the now the technological potential to access to vast sources of satellite imagery to extract environment information for classification, this may not be sufficient to achieve the greatest performance. It is important to also consider the fact that sparse training samples may not be representative of the greater population of data. Testing was conducted on each year of training data and the SMOTE method was modified to achieve an equilibrium of SD classes and provide balance in the sample to eliminate the likelihood of over-fitting. This recurring issue with many labelling problems and classification research disciplines may skew results if not addressed [42]. SMOTE has been applied with many studies in bioinformatics to address imbalanced data issues. This suggests that the imbalanced data problem is a recurring issue in various disciplines of machine learning for classification that this research hopes to address by avoiding imbalanced data with skewed results [43]. Using over-sampling as opposed to under-sampling techniques, the aim is to increase training potential of a sparse dataset by generating a larger pool of data based on the original limited set.

To address issues concerning the problem of limited sample data for training and imbalanced data, the SMOTE method was proposed to increase the number of instances for each year of data ranging from 100 to 1000 instances while also equalizing the class sizes before classifying SD for each new model. The purpose of this is to analyse whether an increase in simulated instance numbers of synthetic data in addition to the equilibrium caveat, can improve balance and performance over original data classification results. To implement the SMOTE Equilibrium approach, the data was processed to equalise the balance of instances in each set, then SMOTE was applied to increase original instance numbers for each year in multiples of 100 ranging from 100 to 1000 instances. Results of these experiments will highlight the possibility of achieving significant increase in accuracy performance for SD prediction with an increase in data instances or whether the difference between a sparse sample will be rendered negligible. The interest of this research is to discover how classification accuracy performance changes with an increased synthetic data pool as opposed to a limited real-world set. If it is the case that a snapshot of data can provide competitive results when compared with a much larger pool of data, then research can be focused on utilising sparse training data rather than spending time on sourcing field survey experts. This will reduce the requirements and cost of specialist equipment needed for collection of training material, saving acquisition time and resources. Initially experiments were conducted using the entire collection of the training sample inclusive of every year as a training set to test on each year in order to assess how well a varied dataset of original could perform in terms of accuracy.

4 Results and discussion

In this section, results are presented of experiments conducted using methods proposed in this research. This includes the Data Imputation approaches to provide value replacement for partially complete training date. Then we have the ITEA methods to compare three- and five-point SD scale using NB, SVM, J48 and MLP classifiers. Finally, SMOTE Equilibrium results from which the most suitable synthetic instance generation for each year are deduced for consideration in future classification and predictions. Analysis is provided on the contrasting results from each year and comparative analysis on benchmark methods.

4.1 Collective data classification results

Table 2 highlights that using the collective data sample for training is not as beneficial as had been expected. The lowest classification accuracy percentage for each classifier has been highlighted in Table 2 to aid with analysis of yearly classification performance. It is clear that the environment samples vary considerably and are not consistently representative of conditions present. This information results in a more onerous task when making accurate future SD predictions. In terms of classifier performance, MLP was highest with an average classification accuracy of 69.18% whereas NB had the lowest average across all years with 54.27%. As for performance year on year, 2003 achieved 79.62% then 2007 having 53.41%. These results provide information regarding individual classifier performance and training sample year prediction potential that may be of benefit when seeking future RS generated environmental data for making SD predictions.

4.2 MRMD analysis

Following the selection of base classifiers used in experiments, the performance of contemporary machine learning techniques were investigated to help with addressing a multi-label classification problem. To this end, the max-relevance-max-distance feature ranking method was applied as proposed in Ref. [33]. This method was used for

Table 2 Collective data vector classification accuracy

	NB%	SVM%	J48%	MLP%	AVG%
Cross-Val	55.16	66.82	65.47	63.68	62.78
2003	59.26	85.19	85.19	88.89	79.63
2005	58.70	84.78	84.78	84.78	78.26
2007	50	54.55	54.55	54.55	53.41
2008	60.87	50	56.52	56.52	55.98
2009	41.67	66.67	66.67	66.67	60.42

feature appraisal on the sparse training sets before applying base algorithms with the addition of LibLinear for further analysis. As Table 3 shows, 2003 and 2005 provides strong classification performance with the most environmentally unstable years of 2007 and 2008 classifying poorly on average. SVM and LibLinear performance both provided strong performance with SVM showing a slight improvement in 2008 results.

4.3 Data imputation

Results from each year of testing are shown with original value removed in the Data Imputation graphs as shown in Fig. 4. It is evident that using the proposed RCC method can provide a more precise value replacement than alternative imputation methods as our results proved on 70% of test cases. It is observable predominantly in years 2003 and 05 that RCC values overlap the original value line making this a viable replacement model for application to future incomplete training data. When compared with the iterative IRMI approach, RCC performed comparatively well although IRMI provided more accurate value replacements on a number of values during sample years 2005 and 2007 so my be worth considering with future imputation studies and experiment comparisons.

4.4 Incremental transductive ensemble approach

ITEA models are appropriately compared on a point to point scale basis where LTA three-point and LTA five-point can be comparatively analysed. Taking this into consideration, it is noticeable from previous experiments in Fig. 6 that the three-point SD scale performs well in the 70% and 80% confidence level but struggles when using 90% confidence whereas the LTA five-point scale does not provide labels as well as the three-point in the 70% and 80% range but performs better when set to 90% confidence level. When using the default Stacking ensemble with the same four algorithms as a benchmark, Fig. 5 shows the LTA having best performance with STA yielding the lowest percentage labelling accuracy which is as expected due to constraints with STA confidence thresholds.

4.5 SMOTE equilibrium application

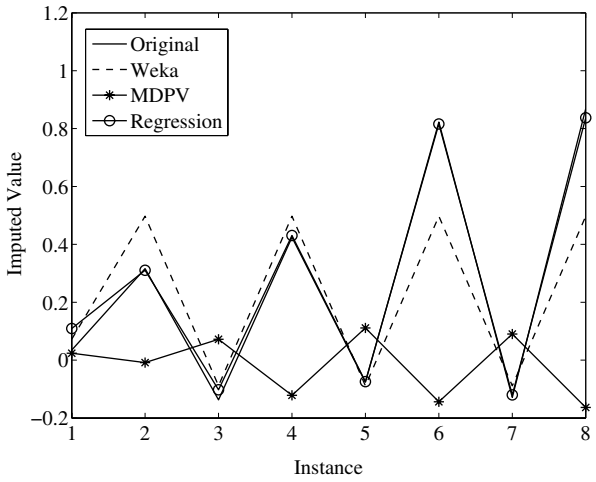
The results of synthetic instance generation simulation testing show a gradual increase in general with some exceptions. For example, in Table 4 it is clear that with 700 and 900 instances there is a reduction in some classification accuracy percentage. This is a matter to investigate further to assess the cause of reduction in these instance brackets in this particular year. As seen in Fig. 7, it is evident when increasing from 1000 to 10000 instances there continues to be good performance with gradual increases made. The exception of this is with NB which seems to level out albeit with high classification accuracy during 2003 and 2005. J48 recorded highest classification performance on average across each year with a general accuracy increase per increment of 1000 instances with the exception of the 2007 sample. This year experienced adverse weather conditions and yielded only partially complete environment training data. Further analysis of these results and investigation into the exploitation/expansion trade-off will reveal how best to generate synthetic instances for addressing this research problem in the future.

Analysis of both Figs. 5 and 6b with STA results and then again with Tables 5 and 6 provided in this paper during the years 2007 and 08, that labelling and classification methods performed poorly when compared with alternative years of data. This is believed to be due to adverse weather conditions and climate issues present over these recorded years rendering less predictable environment feature to SD correlation. It is indicative of these conditions that resulted in partially complete data in year 2007, and also negative correlation in year 2008. If solutions are provided capable of improving classification accuracy and performance consistency during those years which are most difficult to detect SD, then a significant contribution can be made for future disease vector classification.

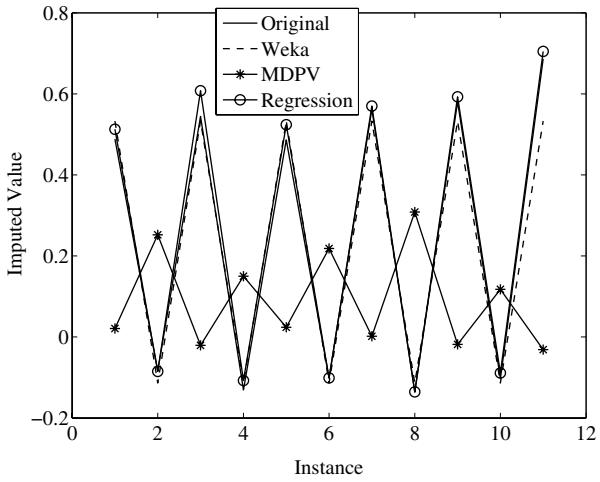
The effect of adverse of weather conditions and flooding from the Yangtze River can have an impact on disease prediction models is a subject requiring further consideration. China and other affected countries at risk of SED can have regular bouts of weather storms and inconsistencies therefore it is vital to provide ways of classifying snail density

Table 3 MRMD results

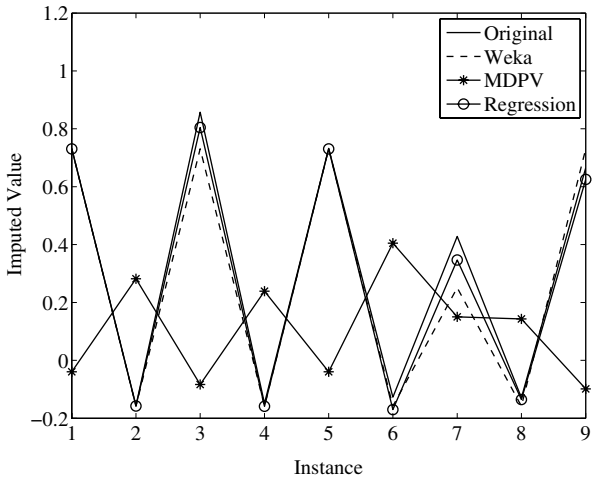
	NB	SVM	J48	MLP	LibLin	AVG%
2003	70.3704	85.1852	81.4815	77.7778	85.1852	80
2005	63.0435	84.7826	71.7391	82.6087	84.7826	77.39
2007	45.4545	54.5455	54.5455	54.5455	45.4545	50.91
2008	56.5217	50	54.3478	56.5217	47.8261	53.04
2009	58.3333	66.6667	63.3333	70	66.6667	65



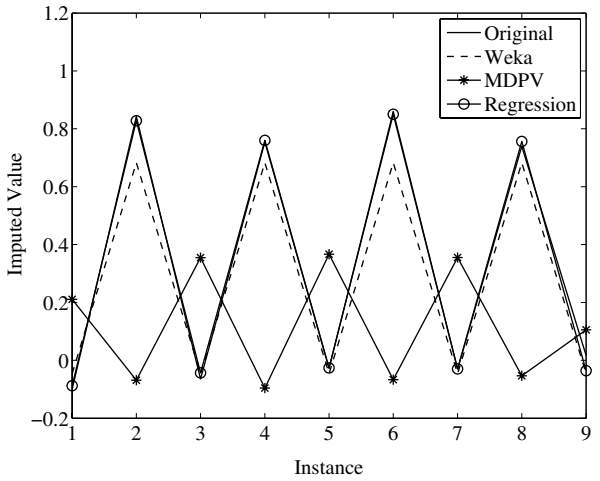
(a) 2003 Data Imputation



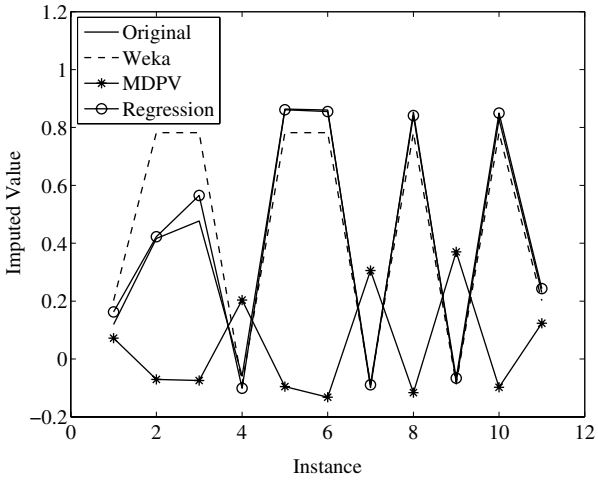
(b) 2005 Data Imputation



(c) 2007 Data Imputation



(d) 2008 Data Imputation



(e) 2009 Data Imputation

Fig. 4 Data imputation methods

Table 4 SMOTE equilibrium results 2005

	NB		J48		SVM	
Instances	Acc. (%)	F-M	Acc. (%)	F-M	Acc. (%)	F-M
100	93.77	0.936	95.44	0.954	79.14	0.759
200	93.88	0.937	97.66	0.977	79.14	0.759
300	93.81	0.937	97.99	0.98	80.43	0.776
400	93.53	0.934	98.44	0.984	83.33	0.815
500	93.63	0.935	98.66	0.987	83.66	0.819
600	94.15	0.94	98.83	0.988	83.63	0.818
700	92.09	0.918	90.65	0.906	48.2	0.422
800	94.24	0.941	90.65	0.906	78.42	0.751
900	93.92	0.938	99.04	0.99	84.09	0.824
1000	94.17	0.94	99.28	0.993	84.89	0.834

Table 5 SMOTE equilibrium results 2007

	NB		J48		SVM	
Instances	Acc. (%)	F-M	Acc. (%)	F-M	Acc. (%)	F-M
100	71.21	0.685	81.82	0.817	39.18	0.35
200	74.24	0.711	87.12	0.871	44.85	0.35
300	73.23	0.699	90.91	0.909	45.7	0.365
400	74.62	0.715	93.56	0.936	42.03	0.42
500	74.96	0.717	96.23	0.962	47.76	0.406
600	74.97	0.718	96.48	0.965	48.05	0.412
700	75.43	0.726	96.75	0.968	50.29	0.44
800	75.76	0.731	96.97	0.97	52.78	0.47
900	75.93	0.733	97.56	0.976	54.55	0.497
1000	76.72	0.742	97.73	0.977	60.69	0.555

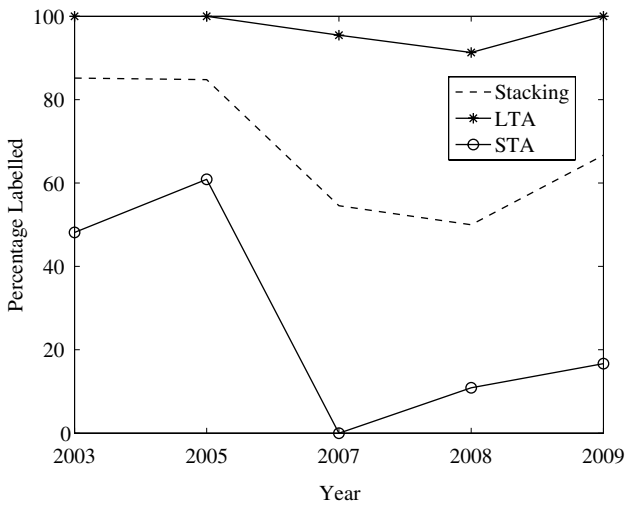


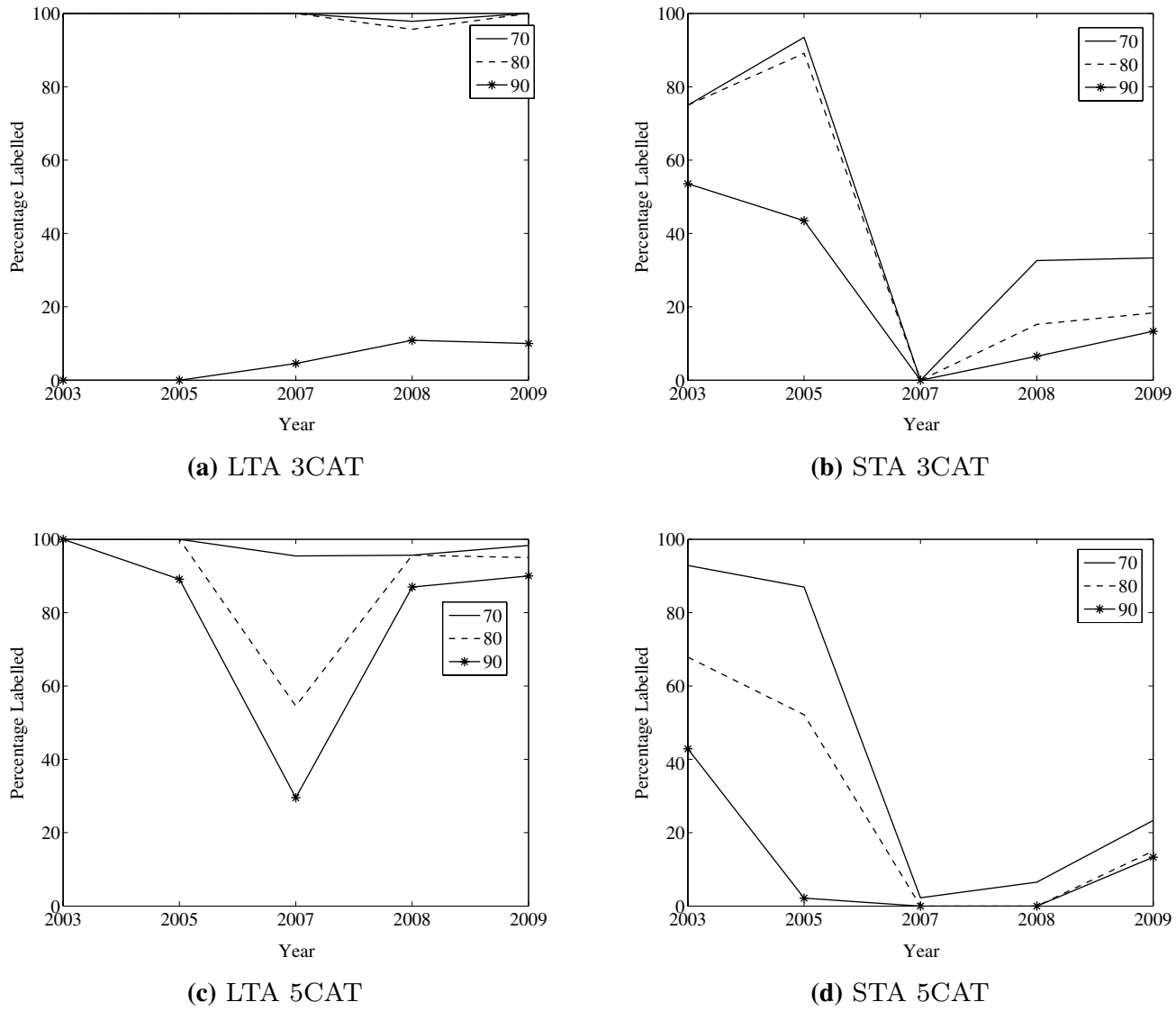
Fig. 5 ITEA comparison

and distribution which can perform well during these conditions. From the proposed solutions for data imputation diagrams in this paper, it is evident that the proposed RCC method outperforms the standard WEKA implementation as well as the previous MDPV approach. In previous experiments [17], MDPV outperformed WEKA by having the lowest percentage difference of replacement compared with the original removed value however, with these new experiments across each year the WEKA method provided better results when compared with MDPV. The novel RCC method has shown to be capable of providing positive results when compared with the previously proposed methods as well as the standard WEKA implementation for value replacement.

4.6 Final results

The results derived from this research are as follows and novel methods are detailed below:

- Proposed RCC method for data imputation has shown to provide superior results in value replacement accuracy when compared with the standard WEKA replacement or previously proposed MDPV methods. Results show a significant increase in value replacement precision using RCC as opposed to other methods as shown in Fig. 4b with more precise values provided on 70% of test cases over other tested methods. This provides the ability to



(a) LTA 3CAT

(b) STA 3CAT

(c) LTA 5CAT

(d) STA 5CAT

Fig. 6 LTA-STA comparative analysis results

improve the training scope of partially complete sample data when faced with future incomplete satellite imagery. IRMI provided interesting results which can be analysed further with more test data.

- When comparing ITSVM with standard SVM application, this research shows that the ITSVM method experiments resulted in greater or equal classification accuracy over standard SVM application in 21% of cases. This is a positive finding which suggests the potential benefits of using incremental, semi-supervised learning for classification in this domain.
- In terms of the ITEA and LTA/STA paradigms, results show that when compared with the standard default Stacking ensemble; LTA outperforms Stacking in each of the including provision of all class labels for 2003, 05 and 09 when using the same four algorithms for experimenting on 29.1%. Figure 6 shows that 70% and 80% confidence have a high labelling percentage whereas it declines steeply when increasing to the 90% level. STA performed inferiorly to the Stacking ensemble in each of the test cases on average of 40.1% of test cases and was incapable of labelling a single instance from the field survey data in 2007 which signifies the extent of classification difficulties in that year. Correlation of labelling percentage decline was evident with both three and five-point category STA results as shown in Fig. 6b labelling performance was especially poor in years 2007 and 2008.
- Each of the SMOTE equilibrium Tables 4, 5, 6, 7 and 8 show that with every increase of 100 instances on average there is a steady improvement in classification accu-

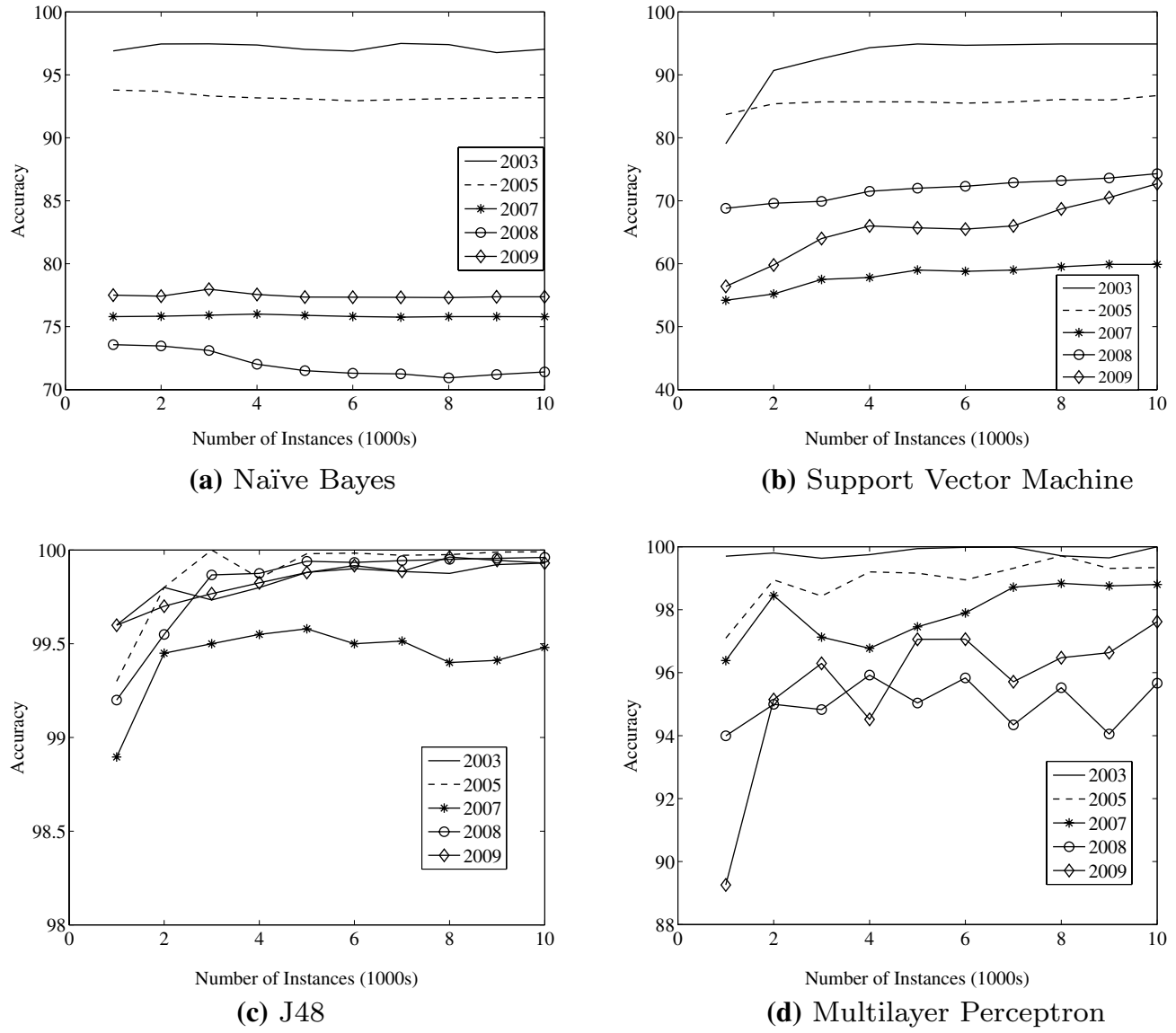


Fig. 7 SMOTE classifier results

Table 6 SMOTE equilibrium results 2008

	NB		J48		SVM	
Instances	Acc. (%)	F-M	Acc. (%)	F-M	Acc. (%)	F-M
100	68.12	0.673	77.54	0.773	70.71	0.691
200	67.39	0.666	88.41	0.885	73.74	0.726
300	67.39	0.666	88.89	0.889	76.1	0.754
400	69.2	0.685	93.48	0.935	75.6	0.748
500	69.28	0.685	94.35	0.943	76.5	0.759
600	71.01	0.704	96.86	0.969	76.24	0.756
700	73.05	0.724	96.6	0.966	77.02	0.764
800	72.79	0.721	97.3	0.973	76.56	0.759
900	72.76	0.721	97.28	0.973	76.11	0.754
1000	71.65	0.709	97.55	0.975	77	0.764

Table 7 SMOTE equilibrium results 2003

	NB		J48		SVM	
Instances	Acc. (%)	F-M	Acc. (%)	F-M	Acc. (%)	F-M
100	89.74	0.897	89.74	0.897	84.31	0.835
200	92.95	0.929	93.59	0.936	92.65	0.925
300	93.59	0.935	94.87	0.949	92	0.919
400	94.87	0.948	96.79	0.968	94.61	0.945
500	95.13	0.951	97.18	0.972	94.51	0.944
600	94.02	0.939	95.51	0.955	95.1	0.95
700	96.52	0.965	97.44	0.974	94.92	0.95
800	96.63	0.966	97.92	0.979	95.45	0.954
900	96.58	0.966	97.86	0.979	95.49	0.954
1000	96.92	0.969	97.82	0.978	95.63	0.956

Table 8 SMOTE equilibrium results 2009

	NB		J48		SVM	
Instances	Acc. (%)	F-M	Acc. (%)	F-M	Acc. (%)	F-M
100	76.67	0.758	86.67	0.865	58.33	0.551
200	76.67	0.757	93.06	0.93	58.06	0.547
300	76.85	0.759	94.81	0.948	57.78	0.546
400	77.55	0.765	96.51	0.965	54.39	0.518
500	77.63	0.766	96.76	0.967	53.8	0.513
600	77.87	0.768	97.95	0.979	57.52	0.544
700	78.58	0.776	98.24	0.982	60.91	0.583
800	79.26	0.783	98.4	0.984	61.87	0.598
900	79.05	0.781	98.34	0.983	61.49	0.594
1000	78.77	0.778	97.95	0.98	61.14	0.589

racy percentage with some individual exceptions. Algorithm classification accuracy % and *f-measure* results are presented in these tables with lowest and highest instance results highlighted to identify range of performance improvement. It is also evident that the SVM classifier performed poorly compared with NB and J48. In accordance with this, the MLP algorithm was applied when testing the SMOTE Equilibrium on the 1000-10,000 instance range in order to diversify the methods further and provide performance analysis. Using synthetic data simulation for each year with instance numbers ranging from 1000-10,000 as shown in Fig. 7 shows a trend in increased performance with all but the NB classifier which remained at consistent accuracy levels throughout each year. Analysis of SMOTE Equilibrium methods show many classification accuracy increases when experimenting from 100 to 1000 instances with an average collective increase of 21.26%. With each increase further from 1000 to 10,000 instances over each year, results indicate a slower collective average accuracy rise of 4.82%.

5 Conclusion

The problem being addressed with this research focuses on issues surrounding labelling snail density and distribution levels for making future predictions of *schistosomiasis* disease likelihood. A number of prediction models have been proposed and implemented to address this issue with key contributions including solutions for data imputation, SD class labelling and sparse training data classification performance improvement. A method summary and key results are detailed below:

- *RCC* to provide a data imputation method that can provide consistent and precise value replacement for incomplete training data acquired from satellite image extraction.
- *Key findings* RCC outperformed standard WEKA value replacement and previously proposed MDPV in over 70% of tests and will now be considered as the most appropriate approach going forward with the research.

- *Incremental transductive ensemble approaches* to enable a robust multi-classifier solution for labelling SD classes. The STA and LTA model caveats provide a bespoke labelling solution that can be applied to a variety of different classification problems.
- *Key findings* The ITEA performance can be difficult to compare on an equal basis with standard ensemble methods however when benchmarking with the default stacking ensemble method, results show significant increase from proposed LTA method in terms of labelling percentage achieved.
- *SMOTE equilibrium* method to assess synthetic instance generation viability and classification performance while avoiding potential over-fitting issues.
- *Key findings* Implementing the SMOTE Equilibrium application has shown in general that classification accuracy results gradually increase in line with synthetic instance generation in the 100–1000 instance range. There are some anomalies to this statement which will be investigated further but in the main course of experiments, improved performance is noticeable with an increase in generated instances. When increasing in the 1000 instance scale range, there is a slight increase in accuracy performance as instance levels increase from 1000 to 10000 instances. This again can be subject to change with some instances, usually from years 2007 to 2008, being problematic in terms of classification performance.

5.1 Technical research contributions

Research contributions of this study include the above methods which include a novel data imputation method (RCC) to address the issue of missing environment data from the real-world sparse training set using a combination of regression and correlation of existing data. ITEA are proposed to improve the snail density labelling potential when dealing with limited training data. This approach was initially applied with the traditional SVM algorithm and then developed further with STA and LTA caveats for suitability and application dependant on the confidence levels required in label providence. SMOTE Equilibrium is introduced as a procedure to provide findings and analysis of classification performance when applying over-sampling techniques to sparse training samples. The data was preprocessed prior to the over-sampling process then comparatively analysed with the original set classification to discover the snail density improvements made with a larger data sample.

Collective data training analysis shows that while some years such as 2003 and 2005 performed well when trained with the entire combined data collection, the remaining years did not fare as well. This is due to inconsistency with environment conditions making it difficult for classification during these alternate years. The training process does not handle the environment anomalies well when classifying so in these cases it may be better to deal with the original smaller dataset for training purposes.

ITEA methods have provided legitimate solutions to the SD class labelling dilemma by using a multi-classifier approach to labelling. Using the two variants of this approach can help utilise the labelling process to suit the research aims. As the results show, LTA provides a high number of labels in the 85% confidence threshold especially in the three-point category graph. The STA as expected labels significantly less instances and shows a steep decline in 2007 and 08 with a small recovery of stability in 2009. These ITEA methods and their results are research problem dependant and can be taken on a problem by problem basis for successfulness of outcome.

SMOTE Equilibrium results show the increase in classification accuracy when adding simulated data instances based on the real-world dataset provided. The difference in accuracy has been detailed through each instance increase from hundreds to thousands to highlight the improvement if any when adding additional instances for training and classification purposes. Results have shown that by using the SMOTE Equilibrium method can significantly increase the sparse data training potential for making future predictions. Additionally, this method prevents over-fitting and enables us to enhance the data synthetically while at the same time keeping the integrity of the real-world sample. These results show that there is a beneficial aspect to the synthetic increase of instances for classification purposes. These results can be further investigated to distinguish the most favourable trade-off between synthetic simulation and real-world data for highly accurate classification for making predictions.

5.2 Future work

Continued research in this domain will include conducting research into how to best approach the changes in environment feature values that derive from poor climate conditions as this will be an issue going forward. Satellite imagery being affected during adverse weather conditions is an ongoing problem that needs to be addressed further to assure the optimal image clarity when applying image extraction for environment features. Looking ahead to the next stage for continuation of this research, consideration will be given to the average temperature of the research areas in these sample years to discover the ratio or relationship of SD values to the increase or decrease in general environment characteristics. Spatio-temporal conditions can then be assessed and proposed methods applied and compared with standard algorithms for classification performance.

Acknowledgements We would like to thank research partners at The European Space Agency and Academy of Opto-Electronics, Chinese Academy of Sciences, China for providing all experiment data used in this work.

Funding Funding was provided by Department for Employment and Learning, Northern Ireland and European Space Agency.

References

1. Kerr J (2003) From space to species: ecological applications for remote sensing. Trends Ecol Evol 18(6):299–305
2. Zscheischler J, Mahecha MD, Harmeling S, Reichstein M (2013) Detection and attribution of large spatio-temporal extreme events in Earth observation data. Ecol Inform 15:66–73
3. Bavia ME, Malone JB, Hale L, Dantas A, Marroni L, Reis R (2001) Use of thermal and vegetation index data from earth observing satellites to evaluate the risk of schistosomiasis in Bahia. Brazil Acta Tropica 79(1):79–85
4. Weng J, Xu Y, Sharma AR (2012) Epidemic analysis and Visualization based on digital earth spatio-temporal framework 2. State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications of Chinese Academy of Sciences, Beijing, 100101, China * Corresponding a, pp 7220–7223
5. Wu Y, Lee G, Fu X, Hung TGG (2008) Detect climatic factors contributing to dengue outbreak based on wavelet, support vector machines and genetic algorithm. Lect Notes Eng Comput Sci
6. Study AC, Poyang OF, Province J (2006) Indicator development for potential presence of schistosomiasis japonicum's vector in lake and marshland regions. Eur Space Agency (Special Publication) 1851:1
7. Ding X, Li X (2011) Monitoring of the water-area variations of Lake Dongting in China with ENVISAT ASAR images. Int J Appl Earth Observ Geoinform 13(6):894–901
8. Palaniyandi M, Anand PH, Maniyosai R (2014) Spatial cognition: a geospatial analysis of vector borne disease transmission and the environment, using remote sensing and GIS. Int J Mosq Res 1(3):39–54
9. Simoonga C, Utzinger J, Brooker S, Vounatsou P, Appleton CC, Stensgaard AS et al (2009) Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. Parasitology 136(13):1683–1693
10. Ying L, Xinle Y, Yuezhi Z, Xiaoyu M, Fei H, Ke Y (2011) Analysis of spatial and temporal characteristics of the epidemic of schistosomiasis in Poyang Lake Region. Procedia Environ Sci 10(Esiat):2760–2768
11. Andrick B, Clark B, Nygaard K, Logar A, Penaloza M, Welch R (1997) Infectious disease and climate change: detecting contributing factors and predicting future outbreaks. Int Geosci Remote Sens Symp (IGARSS) 4:1947–1949. https://doi.org/10.1109/igarss.1997.609159
12. Zhang Z, Ward M, Gao J, Wang Z, Yao B, Zhang T et al (2013) Remote sensing and disease control in China: past, present and future. Parasites Vectors 6(1):11
13. McManus DP, Loukas A (2008) Current status of vaccines for schistosomiasis. Clin Microbiol Rev 21(1):225–242. https://doi.org/10.1128/CMR.00046-07
14. Walz Y, Wegmann M, Dech S, Raso G, Utzinger J (2015) Risk profiling of schistosomiasis using remote sensing: approaches, challenges and outlook. Parasites Vectors 8(1):163
15. Ding Q, Han J, Zhao X, Chen Y (2015) Missing-data classification with the extended full-dimensional Gaussian mixture model: applications to EMG-based motion recognition. IEEE Trans Ind Electron 62(8):4994–5005
16. Karmaker A, Kwek S (2005) Incorporating an EM-approach for handling missing attribute-values in decision tree induction. In: Fifth international conference on hybrid intelligent systems, vol 2005. https://doi.org/10.1109/ICHIS.2005.64
17. Fusco T, Bi Y, Wang H, Browne F (2016) Incremental transductive learning approaches to schistosomiasis vector classification. In: Dragon 4 symposium
18. Fusco T, Bi Y (2016) Medical artificial intelligence modeling (MAIM). In: Iliadis L, Maglogiannis I (eds) A Cumulative training approach to schistosomiasis vector density prediction, vol 475. Unknown host publication, pp 3–13. ISBN 978-92-9221-304-6
19. Hosseinzadeh M, Eftekhari M (2015) Improving rotation forest performance for imbalanced data classification through fuzzy clustering. In: Proceedings of the international symposium on artificial intelligence and signal processing, AISP 2015, pp 35–40. https://doi.org/10.1109/AISP.2015.7123535
20. Li T, Yang J, Chen Z (2010) The early warning and prediction method of flea beetle based on maximum likelihood algorithm ensembles. In: Proceedings of 2010 6th international conference on natural computation, ICNC 2010, vol 4(Icnc), pp 1901–1905
21. Zagorecki A (2014) Feature selection for Naive Bayesian network ensemble using evolutionary algorithms. In: 2014 federated conference on computer science and information systems, pp 381–385. https://doi.org/10.15439/2014F498
22. Bisong HBH, Jianhua GJG (2010) Support vector machine based classification analysis of SARS spatial distribution. In: Natural computation (ICNC), 2010 sixth international conference on, vol 2(Icnc), pp 924–927
23. Ostfeld RS (2009) Climate change and the distribution and intensity of infectious diseases. Ecology 90(4):903–905
24. Elshazly HI, Elkorany AM, Hassanien AE, Azar AT (2013) Ensemble classifiers for biomedical data: performance evaluation. In: 2013 8th international conference on computer engineering systems (ICCES), pp 184–189. https://doi.org/10.1109/ICCES.2013.6707198
25. Iliou T, Anagnostopoulos CN, Stephanakis IM, Anastassopoulos G (2017) A novel data preprocessing method for boosting neural network performance: a case study in osteoporosis prediction. Inf Sci 380:92–100
26. Fathima SA, Hundewale N, Member S (2012) Comparitive analysis of machine learning techniques for classification of arbovirus. In: Proceedings of 2012 IEEE-EMBS international conference on biomedical and health informatics, vol 25(Bhi), pp 376–379
27. Xu H (2006) Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int J Remote Sens 27(14):3025–3033
28. Bruzzone L, Chi M, Marconcini M (2006) A novel transductive SVM for semisupervised classification of remote-sensing images. IEEE Trans Geosci Remote Sens 44(11):3363–3372
29. Köknar-Tezel S, Latecki LJ (2009) Improving SVM classification on imbalanced data sets in distance spaces. In: Proceedings—IEEE international conference on data mining, ICDM, pp 259–267. https://doi.org/10.1109/ICDM.2009.59
30. Quinlan R (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo
31. Salzberg S, Segre A (1994) Review of C4.5: Programs for machine learning by J. Ross Quinlan. Mach Learn 16:235–240. https://doi.org/10.1007/BF00993309
32. Lin C, Chen W, Qiu C, Wu Y, Krishnan S, Zou Q (2014) LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing 123:424–435. https://doi.org/10.1016/j.neucom.2013.08.004

33. Zou Q, Zeng J, Cao L, Ji R (2016) A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing 173:346–354. https://doi.org/10.1016/j.neucom.2014.12.123
34. Zacarias OP, Bostr H (2013) Comparing support vector regression and random forests modeling for predicting malaria incidence in Mozambique. Int J Adv ICT Emerg Reg (ICTer). https://doi.org/10.1109/ICTer.2013.6761181
35. Yang W, Yin X, Xia GS (2015) Learning high-level features for satellite image classification with limited labeled samples. IEEE Trans Geosci Remote Sens 53(8):4472–4482
36. Thomas P (2009) Semi-supervised learning by Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (Review). IEEE Trans Neural Netw 20:542
37. Yu T, Zhang W (2016) Semisupervised multilabel learning with joint dimensionality reduction. IEEE Signal Process Lett 23(6):795–799
38. Ertekin S, Huang J, Bottou L, Giles C (2007) Learning on the border: active learning in imbalanced data classification. In: Proceedings of the international conference on information and knowledge management, pp 127–136. https://doi.org/10.1145/1321440.1321461
39. Kushnir D, Laboratories AlB, Hill M (2014) Active-transductive learning with label-adapted Kernels. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. https://doi.org/10.1145/2623330.2623673
40. Hu S, Liang Y, Ma L, He Y (2009) MSMOTE: improving classification performance when training data is imbalanced. In: International workshop on computer science and engineering, vol 2, pp 3–17. https://doi.org/10.1109/WCSE.2009.756
41. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16(1):321–357
42. Stefanowski J, Wilk S (2008) Selective pre-processing of imbalanced data for improving classification performance. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 5182 LNCS, pp 283–292
43. Malazizi L, Neagu D, Chaudhry Q (2008) Improving imbalanced multidimensional dataset learner performance with artificial data generation: density-based class-boost algorithm. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 5077, pp 165–176
44. Zhang H, Liu G, Chow Tommy WS, Member, Senior and Liu, Wenyin and Member (2011) SeniorTextual and visual content-based anti-phishing: a Bayesian approach. In: IEEE transactions on neural networks, vol 22, pp 1532–1546
45. Wu J-Y, Zhou Y-B, Li L-H, Zheng S-B (2014) Identification of optimum scopes of environmental factors for snails using spatial analysis techniques in Dongting Lake Region, China. Parasites Vectors 7:216. https://doi.org/10.1186/1756-3305-7-216
46. Templ M, Kowarik A, Filzmoser P (2011) Iterative stepwise regression imputation using standard and robust methods. Comput Stat Data Anal 55:2793–2806. https://doi.org/10.1016/j.csda.2011.04.012
47. Hron K, Templ M, Filzmoser P (2019) Imputation of missing values for compositional data using classical and robust methods. Comput Stat Data Anal 54:3095–3107