附录A英文翻译

**原文：**

A predictive model of recreational water quality based on adaptivesynthetic sampling algorithms and machine learning

Tingting Xu[\*](#br1), Giovanni Coco, Martin Neale

School of Environment, Faculty of Science, University of Auckland, New Zealand

a r t i c l e i n f o a b s t r a c t

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Article history:  Received 9 December 2019  Received in revised form  1 April 2020  Accepted 2 April 2020  Available online 13 April 2020  Keywords:  Water quality  Adaptive synthetic sampling algorithm  Nearest neighbour  Boosting decision tree  Support vector machine  Artiﬁcial neural network |  | Predicting recreational water quality is one of the most difﬁcult tasks in water management with major  implications for humans and society. Many data-driven models have been used to predict water quality  indicators to allow a real time assessment of public health risk. This assessment is most commonly based  on Faecal Indicator Bacteria (FIB), with the value of FIB compared with thresholds published in guide-  lines. However, FIB values usually tend to be unbalanced within water quality datasets, with small  proportions of data exceeding guideline thresholds and far larger numbers that do not. This can be a  limiting factor in the uptake of model predictions since, even if the overall accuracy is high, the sensi-  tivity of the predictions can be low. To address this issue, this paper proposes an adaptive synthetic  sampling algorithm (ADASYN) to generate synthetic above-threshold FIB instances and test the validity  of the approach for the prediction of recreational water quality. The models in this paper are based on  four machine learning techniques: k-mean nearest neighbour, boosting decision tree, support vector  machine, and multi-layer perceptron artiﬁcial neural network and are applied to ﬁve different locations  in Auckland, New Zealand. Aside from support vector machine, all models provide favourable predictions  with relatively high sensitivity (around 75%) and overall accuracy (over 90%), indicating that both the  compliant and exceedance conditions can be effectively predicted through the use of more sophisticated  model training which involves artiﬁcial data. Considering the model accuracy and stability, boosting  decision trees (BDT) and multi-layer perceptron artiﬁcial neural (MLP-ANN) network are the best two  models and the multi-layer perceptron is the most efﬁcient with the shortest computation time.  © 2020 Elsevier Ltd. All rights reserved. | | |
| 1. Introduction  Robust and reliable information about the water quality at  beaches is of vital importance as it is used to inform managers  about public health risk from recreation in water ([Thoe and Lee,](#br11)  [2013](#br11); [Stidson et al., 2012](#br11); [Francy et al., 2013](#br10); [Wade et al., 2008](#br11)).  Beach water quality is widely monitored using faecal indicator  bacteria (FIB), which commonly involves the measurement of  enterococci or Escherichia coli concentrations in water. Monitoring  results are usually assessed for compliance with guidelines, and  guideline exceedances trigger advisories or warnings for beach  users ([de Brauwere et al., 2014](#br10); [Castelletti et al., 2014](#br10); [Marshall](#br11)  [et al., 1997](#br11); [King and Richardson, 2003](#br11)). The laboratory tests for  FIB typically take up to 24 h using the most probable method  (MPN), however, once the time for sample collection and  \* Corresponding author.  E-mail address: [txu648@aucklanduni.ac.nz](mailto://(null)txu648@aucklanduni.ac.nz) (T. Xu). | | |  | interpretation and communication of results are included, the test  result is not publicly available probably for up to 48 h after sample  collection. This results in water quality information that is poten-  tially out of date because water quality can change faster than the  analysis time. For this reason, much effort has been invested in  trying to predict FIB concentrations in real time, with a range of  process and data driven based models utilised for this purpose (e.g.  [Bae et al., 2010](#br10); [Abyaneh, 2014](#br10); [Thoe and Lee, 2013](#br11); [He and He,](#br11)  [2008](#br11)).  The vast majority of models used to predict water quality are  data-driven, statistical and categorical, and are therefore sensitive  to the data used in their development ([Thoe et al., 2014](#br11); [Chen and](#br10)  [Liu, 2015](#br10); [Li et al., 2019](#br11)). For example, all these models faced the  issue of unbalanced datasets, whereby the number of data points  that are compliant with guidelines is always much more than the  data exceeding the guideline thresholds for FIB. Therefore, the  models are usually good at predicting compliant conditions rather  exceedance condition. [Chandramouli et al. (2007)](#br10) and [Tufail et al.](#br11) |

<https://doi.org/10.1016/j.watres.2020.115788>

0043-1354/© 2020 Elsevier Ltd. All rights reserved.

2 T. Xu et al. / Water Research 177 (2020) 115788

|  |  |  |
| --- | --- | --- |
| [(2008)](#br11) used artiﬁcial neural networks (ANN) to predict water  quality and found that they slightly outperformed traditional  regression models. However, they only considered the overall ac-  curacy which can lead to a signiﬁcant bias due to the original  dataset which contains a disproportionately large proportion of  compliant data points (i.e., when the FIB is below the guideline  threshold). In [2010](#br11), Kazemi Yazdi and Scholz, assessed water  quality based on surface runoff through an ANN model and  compared the ANN with a multi-linear regression model (MLR).  The ANN outperformed MLR in predicting the runoff treatment  since it has a higher ability to capture the nonlinear relationship  between water quality and microbial factors. They focused on the  exploration of the relation between the quality indicator and the  impact factor but failed to provide details on the model validation  aspect. [Qin et al. (2012)](#br11) introduced an advanced Boosting-tree  based machine learning model but did not compare their results  to other models. [Thoe et al. (2014)](#br11) applied three regression models  as well as classiﬁcation trees (CT) and ANN. They assessed model  results in terms of overall accuracy, but also sensitivity and speci-  ﬁcity, and concluded that CT and ANN performed better than  regression models. Here the overall accuracy refers to the model  successfully predicting below and above threshold data, sensitivity  assesses the model’s ability to predict only exceeding data, and  speciﬁcity assesses the prediction of compliant data. Even in this  case, sensitivity was still low, probably as a result of the bias to the  much more non-exceeding data in the training dataset ([Bedri et al.,](#br10)  [2016](#br10); [Wang et al., 2016](#br11); [Shaw et al., 2017](#br11)). Besides these models,  the capability of multi-layer perceptron-ANN to predict water  quality has been demonstrated in recent years ([Zhang et al., 2015](#br11);  [García-Alba et al., 2019](#br11)). However, such studies considered ANN as  a component to be integrated within other models and did not  present comparisons with other machine learning techniques.  [Granata et al. (2017)](#br11) and [Haghiabi et al. (2018)](#br11) used several ma-  chine learning methods to forecast water quality and reported a  signiﬁcant overestimation of good water quality but did not explore  the reasoning. Moreover, the data sets they used to training the  networks were still unbalanced.  It is well known that ANN and other machine learning methods  need a good quality training dataset to provide reliable and accu-  rate results. In [2012](#br11), Motamarri and Boccelli developed a machine  learning model to classify the recreational water quality at a river  scale. They found that a neural-based learning vector quantization  model outperformed both MLR and ANN, and revealed that all the  predictions heavily relied on the validity of the characteristics of  the training samples. However, the biggest issue for recreational  water quality datasets stay unsloved that they are always unbal-  anced, with the vast majority of the dataset below guideline  thresholds and only a very small amount of data exceeding the  guideline thresholds ([Thoe et al., 2012](#br11); [García-Alba et al., 2019](#br11)).  When training data-driven models, an unbalanced dataset in-  creases the possibility of information loss in the minority class and  overﬁtting in the majority class, which cannot be easily addressed  by only reducing or duplicating data samples ([Batista et al., 2004](#br10)).  [Kim et al. (2014)](#br11) also pointed out to this problem when trying to  monitor the water quality using satellite data using three machine  learning approaches. The limited training data, especially for bad  water conditions, weakened the predictions. To address the un-  balanced sample issue, the Synthetic Minority Over-sampling  Technique (SMOTE) is often used in other ﬁelds ([Han et al., 2005](#br11);  [Luengo et al., 2011](#br11)). SMOTE generates artiﬁcial samples by inter-  polating values between majority and minority classes. However,  using this method, the ‘synthetic’ samples are still more likely to be  distributed in the interior of the minority class ([Luengo et al., 2011](#br11)).  This limitation can be overcome through the advanced Adaptive  Synthetic Sampling algorithm (ADASYN). ADASYN, which can |  | create more samples at the boundary between the two classes and  improve training accuracy ([He et al., 2008](#br11); [Gosain and Sardana,](#br11)  [2017](#br11)). However, in the published ADASYN studies, the data  dimension is usually low with no more than three input variables.  Therefore, its efﬁciency on high dimensional data still needs to be  assessed. In addition, most studies focus on one site and one model,  or at best compare two types of models at one location, usually ANN  and linear regression are compared ([Thoe et al., 2012](#br11); [Chan et al.,](#br10)  [2013](#br10)). [Thoe et al. (2014)](#br11) and [Danades et al. (2016)](#br10) addressed the  issue by comparing three or more different models to reveal their  strengths and weaknesses but only at one location. Other studies  focused on multiple locations, but considered only one model  ([García-Alba et al., 2019](#br11)). Therefore, a sensitivity analysis using  different locations and different algorithms has not yet been  performed.  To overcome the aforementioned deﬁciencies, our study pro-  vides a comprehensive comparison of four machine learning al-  gorithms predicting FIB at ﬁve beaches in Auckland, New Zealand.  The objectives are: (1) to address the problem of unbalanced water  quality datasets by applying ADASYN to improve the training of the  predictive algorithm; (2) to assess the ability of machine learning  models when using a balanced dataset; (3) to compare the different  model results at different locations using different machine  learning models. By completing these objectives, we try to answer  two questions: Can we use artiﬁcial water samples to balance the  original dataset for machine learning models? and does a balanced  dataset improve prediction capability?  2. Study area and data  2.1. Study sites  The ﬁve study sites, Milford Beach, Narrow Neck Beach, Judges  Bay, Weymouth Beach, and Clarks Beach, are distributed around the  coastline of Auckland, New Zealand, and are characterized by  distinct geographical conditions ([Fig. 1](#br3)). The ﬁrst two are open  coast beaches, facing the Paciﬁc Ocean and are wave-dominated.  The remaining three sites are in sheltered bays and harbours, so  are dominated by tidal processes. In 2017, Auckland Council  launched a revised model based beach water quality programme  (‘Safeswim’). The use of models was guided by their ability to meet  performance standards published by the United States Geological  Survey (USGS e [Francy et al., 2013](#br10)). Based on this guidance, a model  with accuracy over 85%, along with sensitivity at 50% and speciﬁcity  at 80%, should be considered as a qualiﬁed model to predict the  water quality at a standard level. For a water quality predictive  model to be included into the safeswim platform, it must meet the  benchmarks in terms of overall accuracy, sensitivity (ability to  predict above threshold) and speciﬁcity (ability to predict below  threshold). Models are in a constant state of review and are reﬁned  regularly in response to Auckland Council’s ongoing beach  sampling.  2.2. Data  Long-term water quality monitoring data is available from 1995  to 2018, with accompanying environmental variables (2019 is also  available for Clarks), for the ﬁve study sites. Each dataset contains  eight variables ([Fig. 2](#br4)): FIB which is used as a threshold to identify  the water quality (single sample threshold ¼ 280, FIB ꢀ 280: ex-  ceedance and FIB < 280: compliance) as prescribed in the New  Zealand Guidelines, rainfall accumulations for the proceeding  twenty four, forty eight, and seventy two hours (one, two, three  days), and the total accumulated precipitation, wind direction and  speed, and solar hours per day. [Fig. 2](#br4) shows that Weymouth has the |

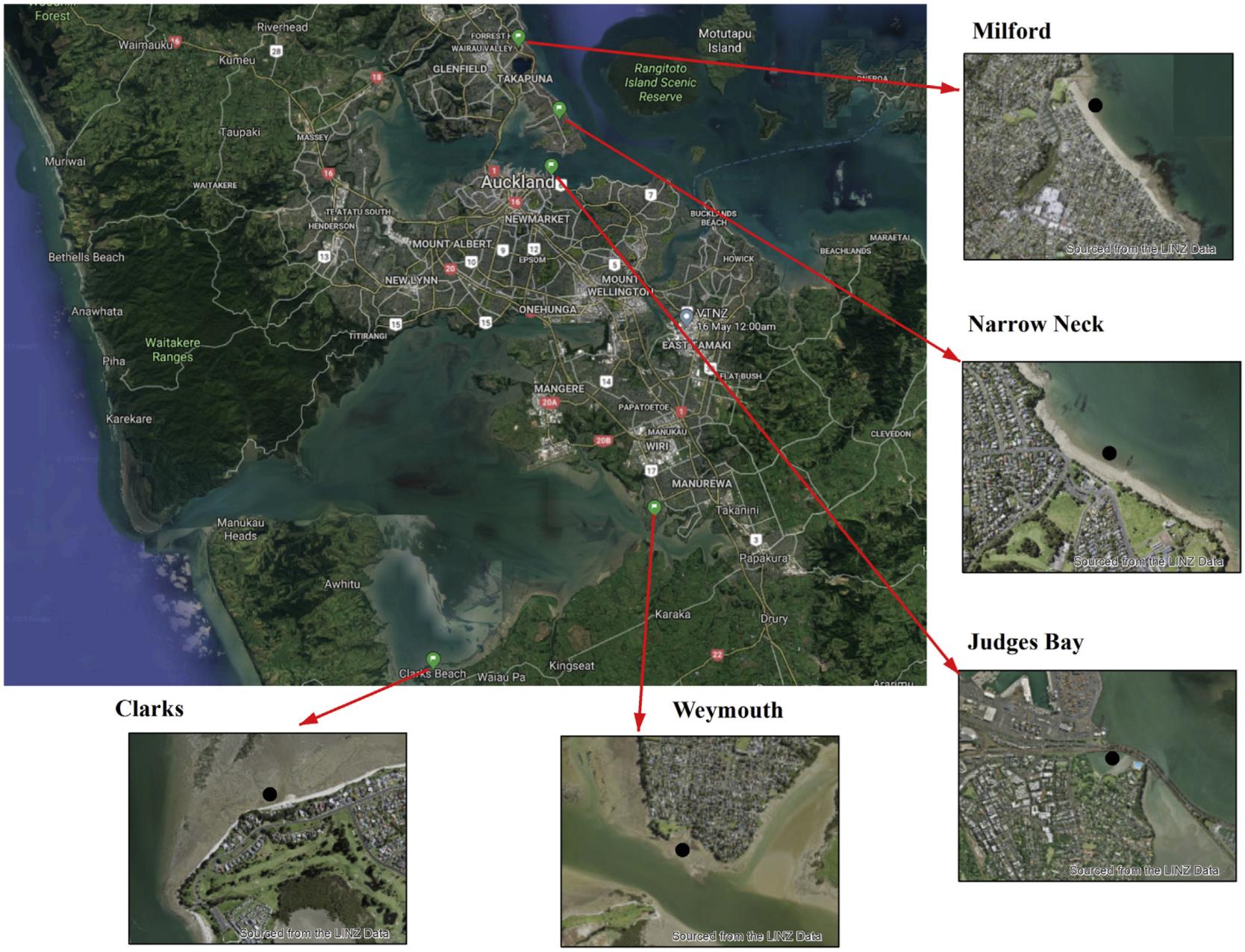
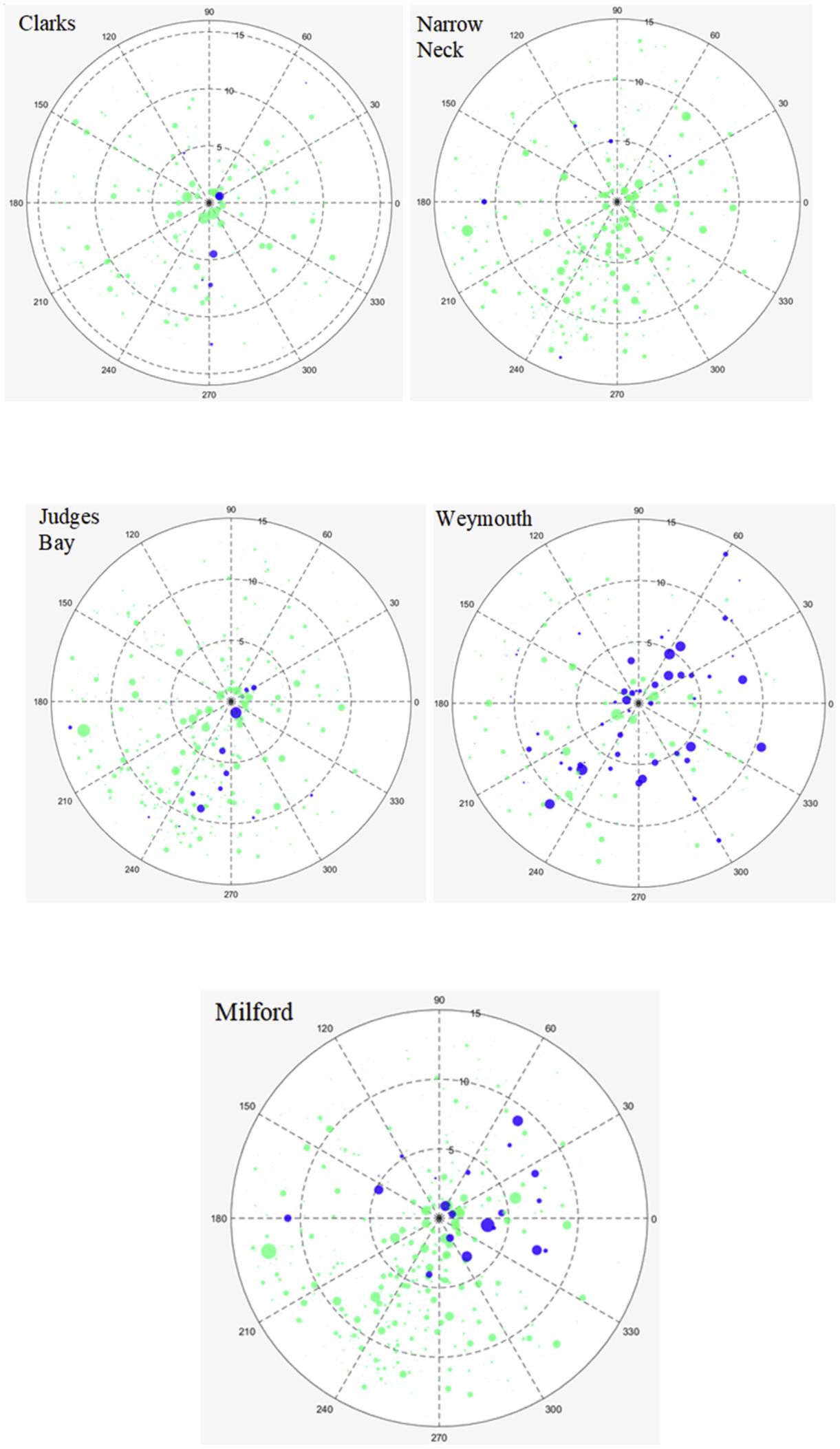
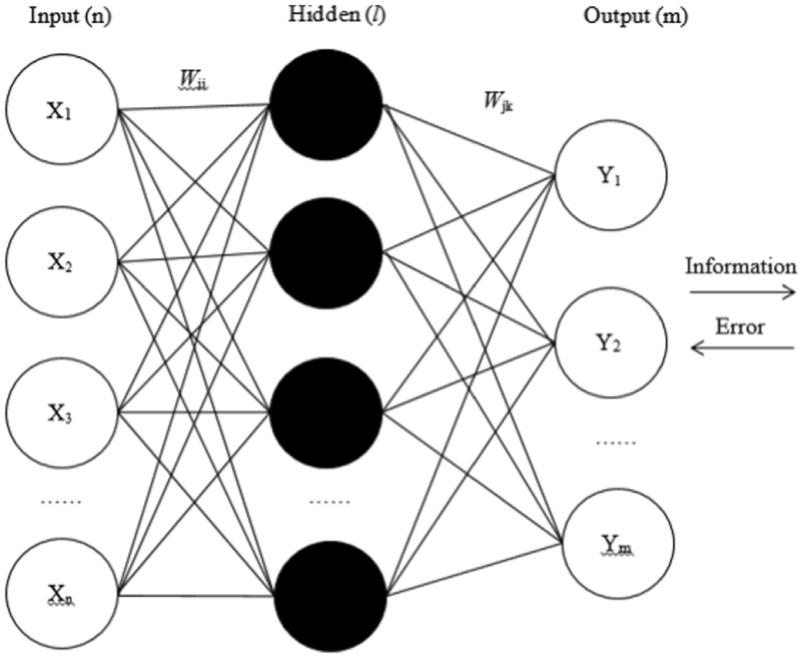
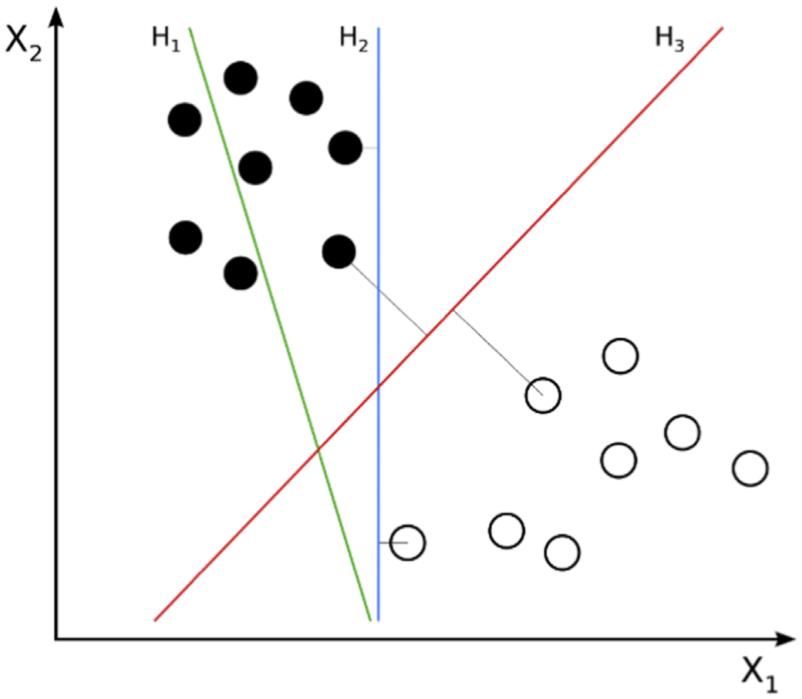
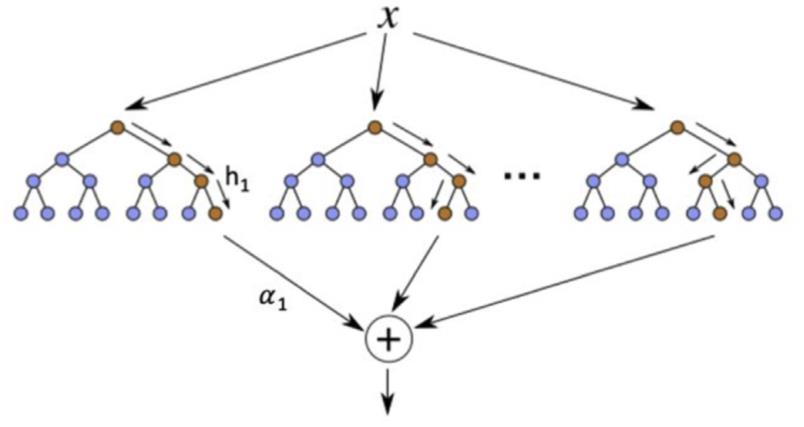
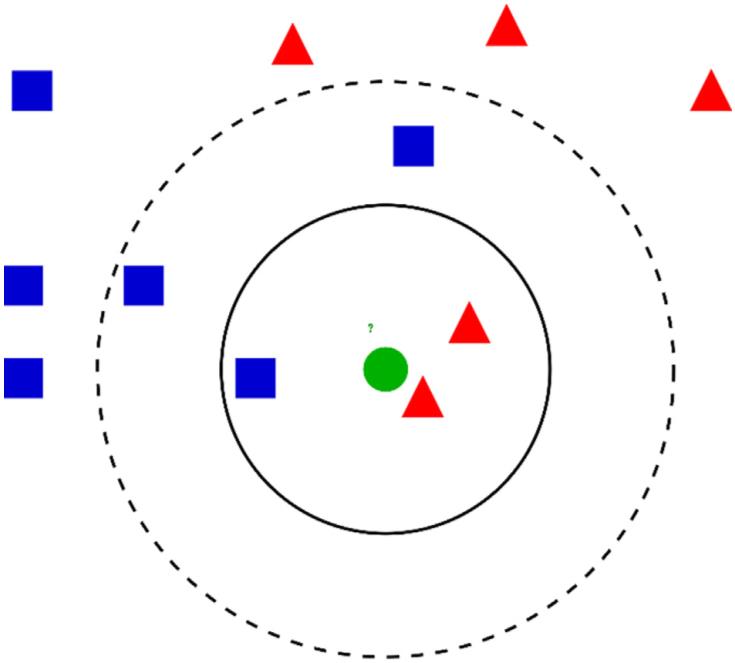
T. Xu et al. / Water Research 177 (2020) 115788 3

Fig. 1. The ﬁve water monitoring sites in Auckland, New Zealand (Black dot in each image indicates the sampling location).

|  |  |  |
| --- | --- | --- |
| most balanced dataset comparing to other four locations, while  Narrow Neck has the most unbalanced dataset, with less than 2.5%  samples found to be above the threshold. The wind direction is  evenly distributed for Clarks and Weymouth, however, the other  three locations experience more southern winds. The average solar  hours of Milford is the smallest with about 6.5 h/day and Judges Bay  has the longest average sunshine time with almost 7 h/day. The  average rainfall accumulation within 72 h in Weymouth is the  greatest, almost 8.2 cm, while it is the smallest at Narrow Neck with  an average rainfall of 5.6 cm in 72 h (see [Fig. 3](#br5)).  3. Methodology  Unlike regression models that usually predict the absolute FIB  value as a continuous function and then reclassify the value based  on threshold, four machine learning algorithms are applied to  ‘predict’ whether the FIB of a water sample has exceeded the  threshold or not through a binary classiﬁcation process.  3.1. K-nearest neighbors algorithm (K-NN) |  | consists of the k closest training examples in the feature space. The  neighbors are taken from a set of objects for which the class (for k-  NN classiﬁcation) or the object property value (for k-NN regression)  is known ([Fig. 3](#br5)). This approach is widely used in classiﬁcation  problems ([Babbar and Babbar, 2017](#br10)).  3.2. Boosting decision trees (BDT)  A decision tree is a ﬂowchart-like structure in which each in-  ternal node represents a “test” on an attribute (e.g. whether a coin  ﬂip comes up heads or tails), each branch represents the outcome  of the test, and each leaf node represents a class label (decision  taken after computing all attributes). The paths from root to leaf  represent classiﬁcation rules. Boosting is a method that combines  many weak learners (trees) into a strong classiﬁer and the Boosting  decision trees ([Fig. 4](#br5)) are popular because of their excellent accu-  racy and fast operation ([Prakash et al., 2018](#br11); [Shoaran et al., 2018](#br11)).  3.3. Support vector machine (SVM) |
| The k-nearest neighbors algorithm is a non-parametric method  used for classiﬁcation and regression. In both cases, the input |  | A Support Vector Machine (SVM) is a discriminative classiﬁer  formally deﬁned by a separating hyperplane. In other words, given  labelled training data, the algorithm outputs an optimal hyperplane |

 Fig. 2. Datasets at 5 locations in Auckland. Using a polar coordinate system, the direction represents wind direction (0e360ꢁ), distance from the origin point represents the dailysolar hours (0e15 h), the size of the dot reﬂects the total precipitation amount accumulating within 72 h (0e100 cm), and the colour indicates whether the water quality exceeded the threshold or not (blue: FIB ꢀ 280 and green: FIB < 280). (For interpretation of the references to colour in this

ﬁgure legend, the reader is referred to the Web version of this article.)

T. Xu et al. / Water Research 177 (2020) 115788 5

|  |  |  |
| --- | --- | --- |
| Fig. 3. Example of k-NN classiﬁcation. The test sample (green dot) should be classiﬁed  either as a blue squares or a red triangles. If k ¼ 3 (solid line circle) it is assigned to the  red triangle population because there are 2 triangles and only 1 square inside the inner  circle. If k ¼ 5 (dashed line circle) it is assigned to the blue square population (3 squares  vs. 2 triangles inside the outer circle). (For interpretation of the references to colour in  this ﬁgure legend, the reader is referred to the Web version of this article.) |  | Fig. 5. An example of SVM classiﬁcation. H1 does not separate the classes. H2 does, but  only with a small margin. H3 separates them with the maximum margin. |

Fig. 4. Boosting decision tree: h is the weak classiﬁer and a represents the extent weight assigned to h.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| which categorizes new examples. In two dimensional space this  hyperplane is a line dividing a plane in two parts where each class  lays on a different side ([Fig. 5](#br5)). | | |  | Fig. 6. Architecture of a BP-TLP ANN |
| 3.4. Artiﬁcial neural network (ANN)  ANNs are a widely used modelling technique with self-adapting,  self-organizing, and self-learning abilities ([Pijanowski et al., 2002](#br11),  [Anctil et al., 2004](#br10)). ANNs include an input layer (where variables  are inputted into the algorithm), one or more hidden layers (where  input variables are combined) and an output layer (the prediction).  Because of its simplicity, ease of training, and its ability for  reasonable associative memory and prediction ([Rumelhart et al.,](#br11)  [1986](#br11)), we used a feed-forward, error Back-Propagation Three-  Layer Perceptron (BP-TLP) ANN architecture ([Fig. 6](#br5)). The most sig-  niﬁcant character of a BP-TLP is that during the training stage, the  information is transiting forward while the error is back propa-  gating. During the feed-forward transition, the information (values/  features) of the input layer is processed by the hidden layer(s) and  then transitions back to the output layer. Every node in a layer |  | affects the node state in the following layer(s). If the value of the  output layer does not meet the expected outcome, it turns to the  back-propagation process. Depending on the expected error, the  BP-TLP ANN begins to adjust the weights and threshold values of  the network, allowing for the predicted outcome to approach the  expected result.  3.5. Adaptive synthetic sampling algorithm (ADASYN)  ADASYN is an improved version of the Synthetic Minority Over-  sampling Technique (SMOTE), which is used to avoid overﬁtting  occurring when exact replicas of minority instances are added to  the main dataset ([Gosain and Sardana, 2017](#br11)). The key idea of the  ADASYN algorithm is to use the density distribution as a criterion to  automatically determine the appropriate number of synthetic  samples that need to be generated for each minority data example. | | | |

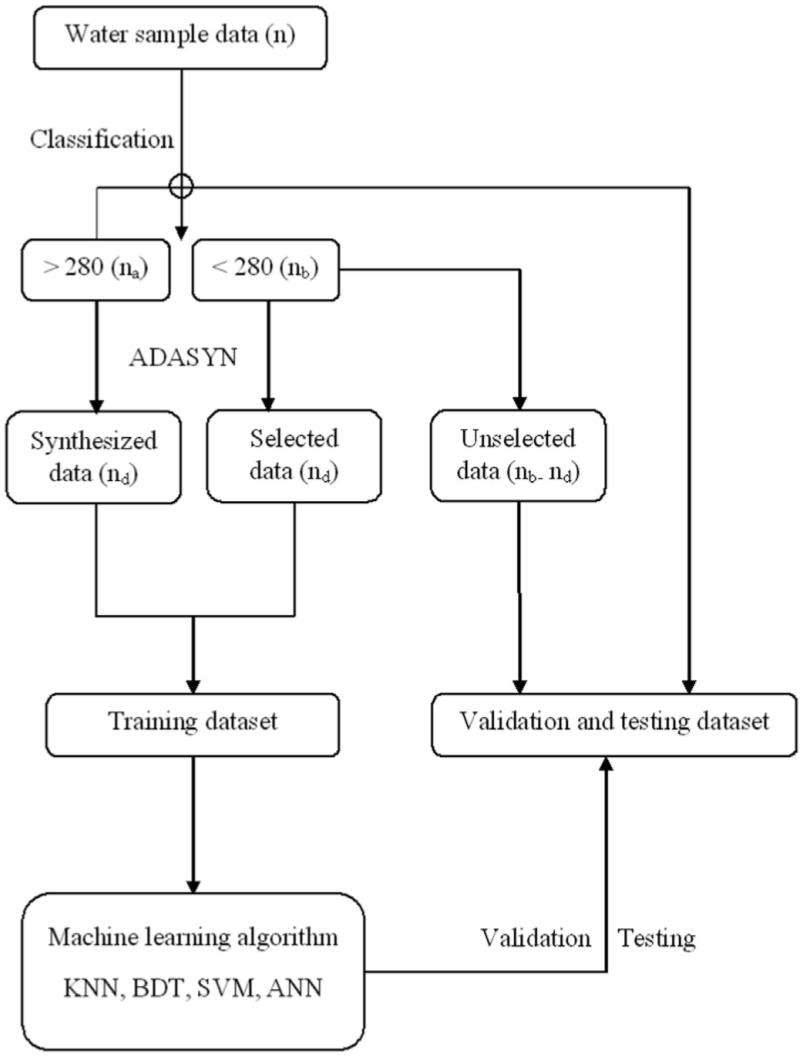
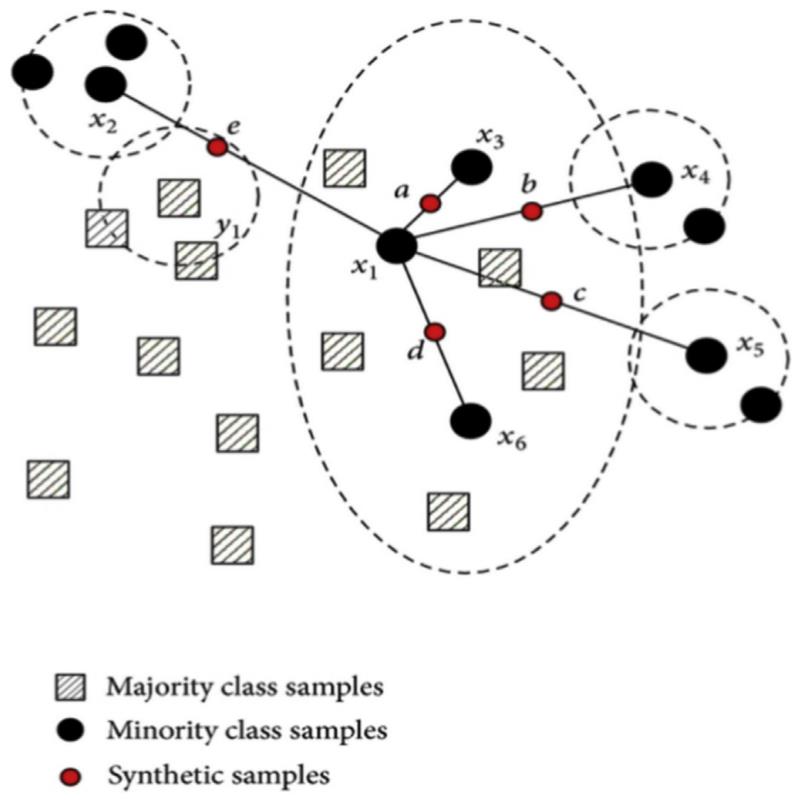
6 T. Xu et al. / Water Research 177 (2020) 115788

Fig. 7. Generation of Synthetic Instances with the help of SMOTE. a-e are the synthetic samples, x1-x6 are the monitory samples.

|  |  |  |
| --- | --- | --- |
| The density distribution can be obtained from the knn function  based on an n-dimensional vector Euclidean distance between  majority and minority samples. A subset of data is taken from the  minority class as an example and then new synthetic similar in-  stances are created ([Fig. 7](#br6)). SMOTE stops here but with ADASYN, a  random small value is added to the ‘faked’ samples making the new  sample more realistic instead of being linearly correlated to the  original sample (more details on the ADASYN application proced-  ures can be found in Appendix I).  3.6. Modelling framework  [Fig. 8](#br6) illustrates the workﬂow we used, showing how the data  are processed and applied to the models. The sampled data makes  the original database, containing two parts as seven training ﬁelds  and one label ﬁeld (FIB). Using FIB values, the samples are classiﬁed  into two classes (FIB above or below 280) to maintain consistency  with the New Zealand guidelines. ADASYN is then applied to  generate synthetic samples above the threshold to balance the  dataset. To form the training dataset, data samples are randomly  selected from the samples below the threshold, 280, and combined  with the synthesised samples. The new training sample dataset is  then fed as input data into the four machine learning algorithms as  input data. The unselected remaining samples below the threshold  (<280) are coupled with the original (not generated through  ADASYN) over-threshold samples (>280) to form the validation and  testing dataset and are applied to the validation process of the  model results.  The four classiﬁers, k-nearest neighbourhood (KNN), boosting  decision tree (BDT), support vector machine (SVM), and multilayer  perceptron (MLP-ANN) were implemented with Matlab. The ﬁrst  three used optimized hyper-parameters with Bayesian optimiza-  tion process to ﬁnd the best parameters. The Bayesian optimization  is a built-in function of Matlab that can optimize most critical pa-  rameters (hyperparameter) automatically. The optimization pro-  cess can minimize the cross-validation loss by varying the |  | Fig. 8. Flow chart of the modelling framework. In parenthesis is the number of sam-  ples. Notice that: n ¼ na þ nb; nd ¼ nb - na.  parameter iteratively. For each individual machine learning algo-  rithm, the critical parameters are different. [Table 1](#br7) describes the  critical parameters and the optimization results.  While MLP-ANN used different method to decide the parame-  ters. First, the number of hidden nodes signiﬁcantly affects ANN  performance. Too few nodes will cause a signiﬁcant prediction er-  ror, while too many will prolong the training process and lead to  overﬁtting. Based on the rule that the number of hidden neuron  should not be less than 2n/3 þ 1 (n ¼ the number of input nodes),  we tried the number 6, 7, 8, 9, and the optimal number of hidden  neurons, 8, was our ﬁnal choice in terms of both the model per-  formance and the network simplicity. Then, the Levenberg-  Marquardt algorithm function was implemented in Matlab to  train the network that updates weights and biases due to its  powerful computation with relatively small datasets and the high  efﬁciency for backpropagation. In addition, the transfer functions  for hidden layer and output layer are a sigmoid (tansig) and a linear  function (purelin), respectively.  Each model was run 100 times for each study site to assess the  model capability and stability. A confusion matrix is used to  calculate the overall accuracy, sensitivity and speciﬁcity, of each  model. The running time is also recorded to evaluate the compu-  tation efﬁciency.  Overall, as shown in [Fig. 8](#br6), for a data D with n samples, na and nb  are the number of water samples above and below the FIB  threshold, where na plus nb equals n. Notice that na is much smaller  than nb and their difference is nd. In our study, the total demand  number of synthesised samples also equals nd. Then, we randomly  selected below-threshold samples from nb and combine them with |

T. Xu et al. / Water Research 177 (2020) 115788 7

Table 1

Hyperparameters be optimized through Bayesian optimization.

Method Name Hyperparameter Description Result (e.g. Clark)

KNN Distance metric Distance searching method of nearest neighbors mahalanobis

Neighbour Number of nearest neighbors to ﬁnd for classifying each point 8

BDT Boosting method Ensemble aggregation method, specially focusing on different boosting functions GentleBoost Number of learning cycle Total number of learns been trained 43 Learning rate The shrinkage range of the learning process 0.0099 Minimum leaf size Minimum number of leaf node observations 81SVM Box constrain The maximum penalty imposed on margin-violating observations 0.0089 Kernel scale A value used to divide all elements of the predictor matrix 0.001

|  |  |  |
| --- | --- | --- |
| an equal amount of the synthesised samples to make the new  training water sample dataset (ntrain ¼ 2nd). The original above-  threshold samples combined with the remaining below samples  (equal to nb - nd), are used for model validation. Before applying the  modelling framework to the ﬁve locations’, a simpliﬁed under and  over sampling method was also applied to reduce the number of  below samples and equal the number of below and above samples.  4. Result and discussion |  | Table 3  Predictions using the raw dataset.  Site Machine Learning  Test Acc. Above Below  Clarks 0.93 0/5 68/68  Narrow Neck 0.94 0/8 135/135  Judge Bay 0.94 0/8 132/132  Weymouth 0.87 0/11 73/73  Milford 0.99 0/2 162/162 |
| 4.1. Original dataset  [Table 2](#br7) describes the water quality monitoring data samples and  [Table 3](#br7) reveals the training results from all the algorithms using the  original dataset. [Table 2](#br7) shows that all the monitoring sites have a  quite unbalanced water samples and only Weymouth has a slightly  better balances the number of samples above and below the  threshold. According to [Table 3](#br7), even though the overall accuracy is  high, none of the 5 locations can be well modelled with the four ML  methods since none of the above threshold samples (exceedances)  were successfully predicted. This high model overall accuracy can  be problematic because of the unbalance dataset issue. Since all the  datasets have very limited occurrences of above threshold samples,  algorithms only focus on the below threshold samples. The high  overall accuracy of each model is due to the correct prediction of  below threshold samples (compliance) while the above samples are  not well predicted. Hence, all of these four models can precisely  predict the below samples but none of the algorithms can identify  the water samples with FIB above the 280 threshold.  4.2. Under/over sampling methods  The under sampling method is ﬁrst used to balance the data  before applying the ADASYN. We reduced the number of below  samples through a random selection process until it equalled the  number of the above samples. Based on this under sampling  dataset, the results [(Table 4](#br8)) indicate that the four aforementioned  algorithms can predict the water quality more accurately. KNN and  BDT have the same results for all the four locations and predict  sample groups from Clarks with a 0.75 overall accuracy. SVM and  ANN are also useful to predict the water sample groups from  Weymouth and Milford, with an overall accuracy at 0.73 and 0.75.  These results make much more sense than using the original data |  | since both the above and below samples can be predicted. However,  other models signiﬁcantly underestimate the below samples at  different locations. Taking Milford for example, the SVM under-  estimated three above samples while overestimated two below  samples.  The underestimation of below samples is due to the loss of  potentially useful information with the under sampling method.  The training sample may be insufﬁcient as well as it may not be a  representative example of the samples below the threshold.  Thereby, even the overall accuracy is acceptable, the result is  inaccurate and inadequate to be used to predict the water quality.  The over sampling method is the opposite of under sampling.  For every location, we duplicated the above samples and then  combined them with the below samples to make a new dataset. The  results of the over-sampling prediction are shown in [Table 5](#br8).  In [Table 5](#br8) we can see that BDT outperformed the other three  algorithms in terms of the best overall accuracy, while SVM per-  formed the worst. BDT, KNN, and ANN all display relatively high  modelling accuracies at all locations, except for Weymouth. How-  ever, results are still deceiving because only replicating the same  above samples with the unchanged sample values can increase the  likelihood of overﬁtting. In addition, if there is some error in the  samples above threshold, the over sampling method will simply  amplify the bias.  4.3. ADASYN approach  Adaptive Synthetic Sampling algorithm (ADASYN) was used to  improve data balance. ADASYN synthetically creates new samples  from the above samples via linear interpolation. This approach  created more samples in the vicinity of the boundary between the  two classes than in the interior of the above samples. This approach | |

Table 2

Water quality samples from the ﬁve monitoring sites (above/below threshold).

Threshold ¼ 280 Clarks Narrow Neck Judge Bay Weymouth Milford

ꢀ280 14 15 23 74 28<280 231 463 444 207 520 Total 245 478 467 281 548Ratio (above) 0.0571 0.0314 0.0493 0.2633 0.0511

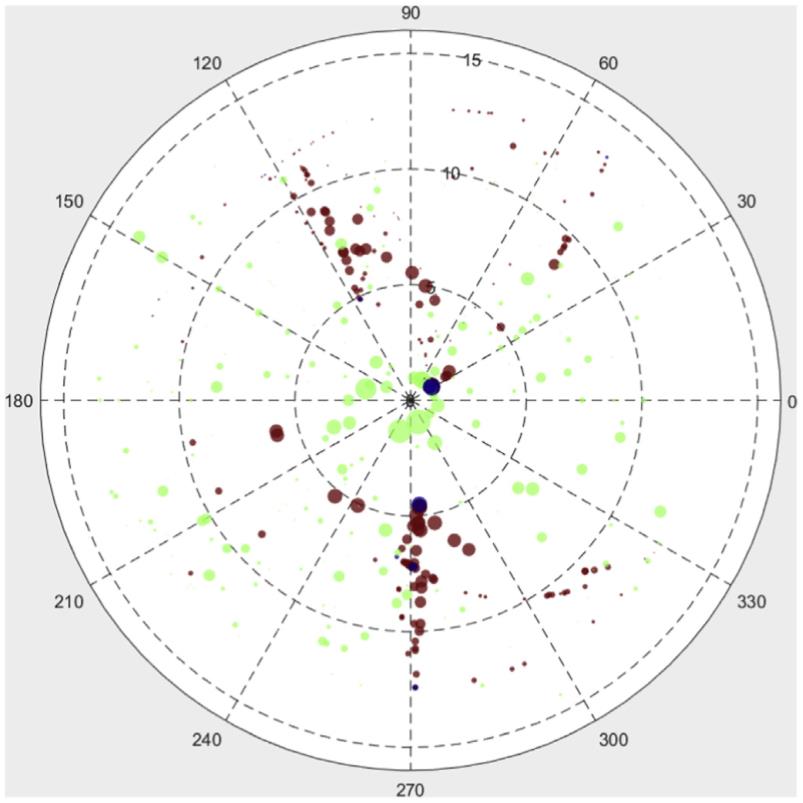
8 T. Xu et al. / Water Research 177 (2020) 115788

Table 4

Under-sampling results.

Site KNN and BDT SVM ANN Acc. Above Below Acc. Above Below Acc. Above Below

Clarks 0.75 3/3 3/5 0.50 3/3 1/5 0.38 3/3 0/5Narrow Neck 0.56 3/3 2/5 0.33 3/3 0/6 0.33 3/3 0/6Judge Bay 0.50 6/8 1/6 0.57 8/8 0/6 0.50 6/8 1/6Weymouth 0.52 11/12 12/32 0.73 0/12 32/32 0.43 9/12 10/32Milford 0.44 6/7 1/9 0.69 4/7 7/9 0.75 7/7 5/9

|  |  |  |
| --- | --- | --- |
| mitigates the problem of overﬁtting caused by duplicated over-  sampling as synthetic new samples are generated artiﬁcially so  that there is no loss of useful information and less risk of over-  ﬁtting. [Fig. 9](#br8) illustrates the dataset obtained with the ADASYN  approach.  We ﬁrst applied the four ML algorithms using the ADASYN  dataset and validated and tested with the original dataset for Clarks  Bay. The results signiﬁcantly improve, see [Fig. 10](#br9) (we show results  using MLP ANN as an example, results for other algorithms can be  found in Appendix II). In [Fig. 10](#br9) (1, 1) and (0, 0) represent how many  water samples are correctly predicted/classiﬁed. The ﬁrst value  indicates the correct prediction for above threshold water samples  (sensitivity) and the second value mirrors the accuracy of below  sample modelling (speciﬁcity)., Overestimation and underestima-  tion are indicated with (0, 1) and (1, 0). The above and below  threshold samples are both precisely predicted by KNN using the  ADASYN dataset for training and the original dataset for validation.  The overall accuracy is 88.43%. The accuracy of speciﬁcity (below  samples) and sensitivity (above samples) is 89.25% and 76.47%,  respectively, both are higher than the under sample methods and  other models. With this balanced dataset, the underestimation and  overestimation are also acceptable since only 10.75% of the above  threshold samples were missed and 23.53% of below samples were  false alarmed to be above. Compared to other studies, the speci-  ﬁcity accuracy does not increase signiﬁcantly but the sensitivity  prediction accuracy improved to over 70%.  The difference between over-sampling and ADASYN is the data  sample itself. For over-sampling, we simply duplicate the existing  minority class samples (above 280) until the number equals to the  majority class (below 280). For example, assume there are 200  samples below the threshold value, 280, and only 20 samples with  a value greater than 280. The simpliﬁed over-sampling will only  copy these 20 above samples for 10 times until reaching the same  number as the below samples and these duplicated samples will  feed into the training and validation process. Therefore, the ﬁnal  model validation result could be really high simply because of using  the same data samples in both training and validation process. In  addition, this approach magniﬁes overﬁtting.  However, with the ADASYN method, the dataset will be  balanced with brand new artiﬁcial above-threshold data samples  by producing new values rather than simply replicating the existing  ones. There is no conﬂict between training and validation dataset. |  | Fig. 9. Balanced data for Clarks Beach, green dot: below sample, blue dot: above  sample, brown dot: ADASYN sample. Direction, distance from the center, and size of  the dots represent the wind direction (0e360ꢁ), daily solar hours (0e15 h), and the  total precipitation amount accumulating within 72 h (0e100 cm). (For interpretation  of the references to colour in this ﬁgure legend, the reader is referred to the Web  version of this article.)  The synthetic samples will only be used to train the model and the  true data will only be applied to the model validation.  Based on above explanations, even if the accuracy value of over-  sampling is higher than ADASYN, we cannot say that the over-  sampling model outperforms the ADASYN due to the data dupli-  cation issue. In contrast, the results from ADASYN are more reliable  because the dataset is balanced and separates the training and  validation datasets by creating artiﬁcial data. |

Table 5

Over-sampling results.

Site KNN BDT SVM ANN Acc. Above Below Acc. Above Below Acc. Above Below Acc. Above Below

Clarks 0.98 72/72 66/69 0.99 64/64 76/77 0.69 46/68 51/73 0.94 59/59 43/49 Narrow Neck 0.98 147/147 126/132 0.99 142/142 136/137 0.68 105/142 85/137 0.97 102/102 105/112 Judge Bay 0.96 127/127 128/138 0.99 132/132 131/133 0.67 64/131 114/134 0.91 80/80 104/123 Weymouth 0.83 65/68 42/61 0.88 62/68 51/61 0.64 47/73 36/56 0.73 31/47 41/52Milford 0.98 154/154 151/157 0.99 163/163 145/148 0.73 93/151 135/160 0.97 109/109 123/130

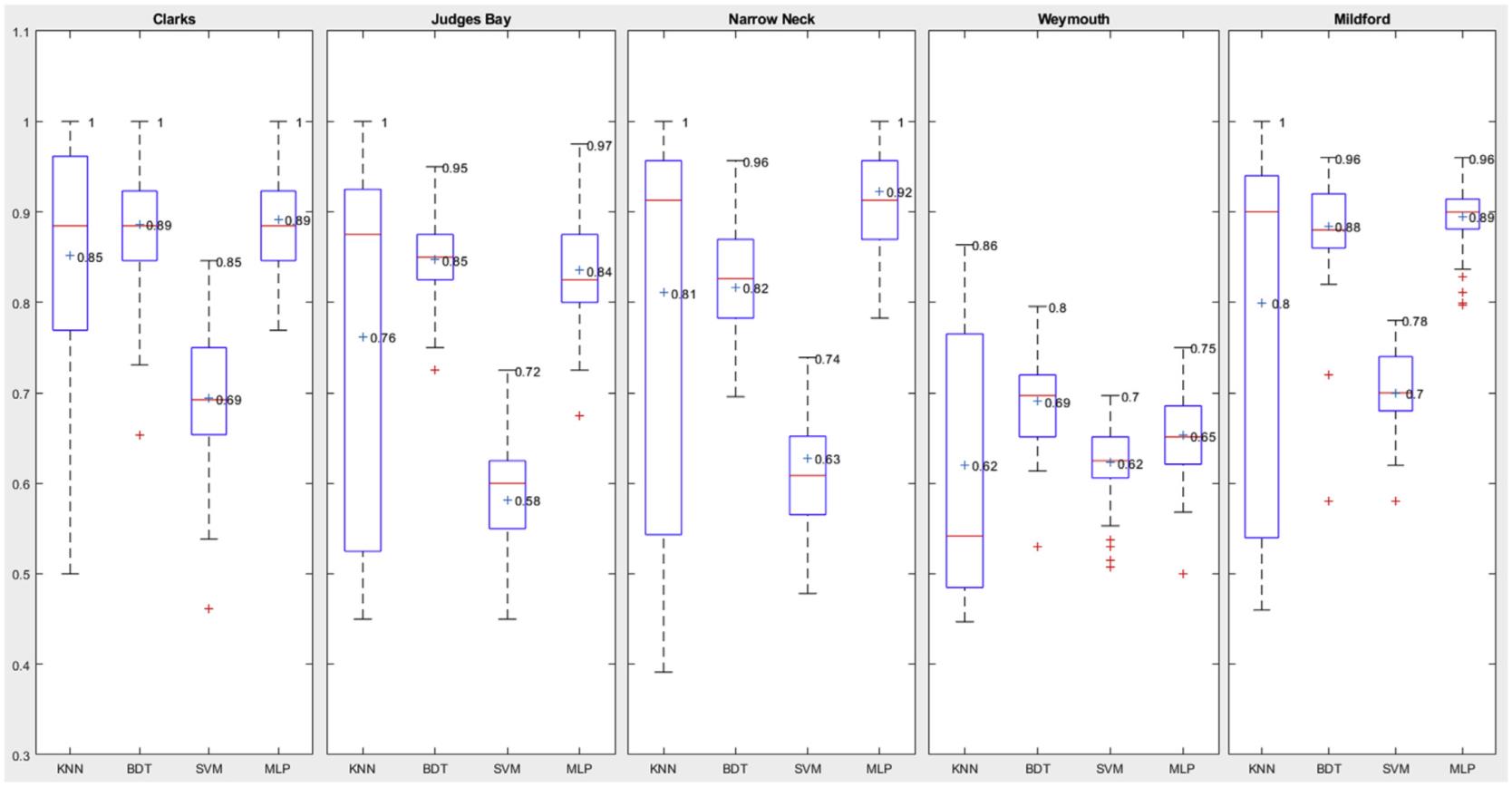
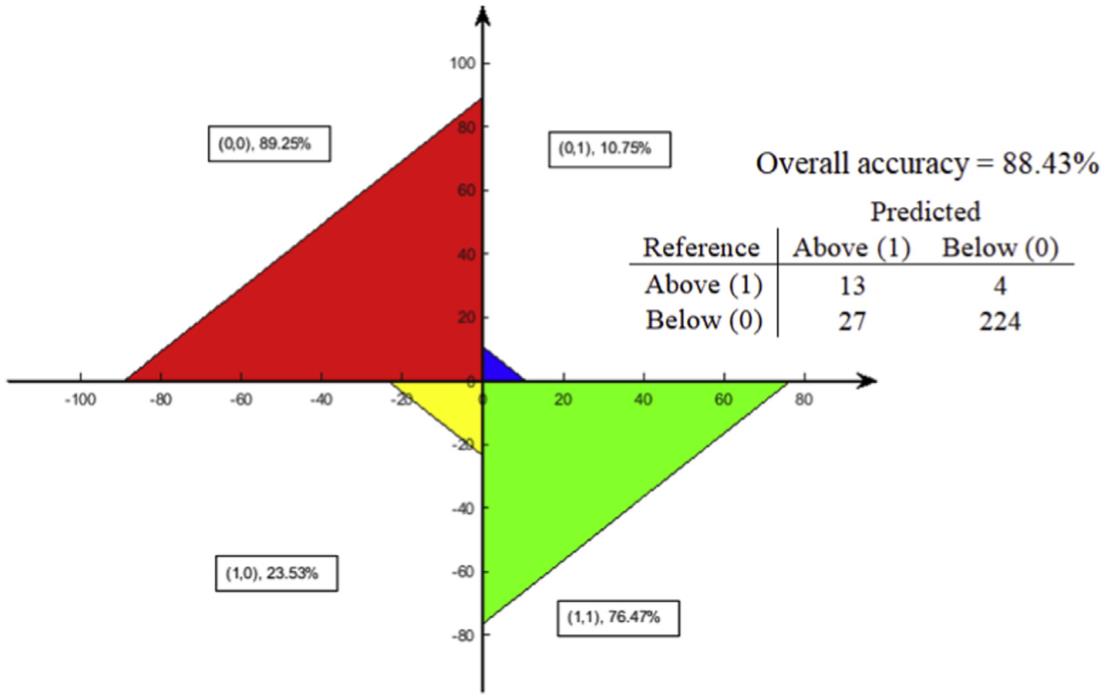
T. Xu et al. / Water Research 177 (2020) 115788 9

Fig. 10. Visualization of the MLP-ANN result using ADASYN, Clarks Beach. Results for other algorithms are presented in Appendix II.

|  |  |  |
| --- | --- | --- |
| 4.4. Comparison four model results among the ﬁve locations  [Fig. 11](#br9) (at the end of this article) displays the validation results  between models and original data for all the locations using four  ML algorithms with ADASYN dataset. Boxplot diagrams, used to  reveal the model accuracy (we performed 100 runs with each al-  gorithm), show the highest accuracy (top line), the mean accuracy  (middle line), the lowest accuracy (bottom line), and the accuracy  range (box size). The red crosses indicate any eventual outlier in the  model predictions.  KNN precisely predicted the above and below samples for all the |  | beaches with an average accuracy around 80%. Except for Wey-  mouth, at least one KNN can predict the above and below samples  correctly, nearly reaching a 100% accuracy. However, the box size of  KNN is always larger than the other three models (the lowest ac-  curacy for each site is always associated to KNN), indicating that a  large variation existed for the 100 runs of KNN. We conclude that  the results are not as robust and reliable as the other models. SVM is  the algorithm showing the worst performance with very low mean  model accuracy values, never higher than 0.7. This model should  not be considered as an appropriate prediction model because of its  inferior performance at all the locations (some of the SVM based |

Fig. 11. Boxplots showing the modelling accuracy at the ﬁve monitoring sites.

10 T. Xu et al. / Water Research 177 (2020) 115788

|  |  |  |
| --- | --- | --- |
| predictions are below 0.5). BDT and MLP-ANN predictions are less  variable than KNN and SVM (relatively smaller box sizes in [Fig. 11](#br9)).  The results from BDT and MLP-ANN are very similar for Clarks,  Weymouth, and Milford. BDT outperformed ANN slightly at Judges  Bay (smaller box size, higher mean, 0.85, and minimum accuracy  values, 0.70), however, the ANN provides a better result at Narrow  Neck with a much higher mean accuracy value (0.92) than BDT  (0.82).  At different location, model performance is also different. Both  BDT and MLP can be used to predict the above and below water  samples for Clarks Beach (high average accuracy, 0.89, and stability,  small box size). For Judges Bay, BDT should be considered as the  most appropriate model because of the second highest mean ac-  curacy value (0.85) and least variation among all the models. For  Narrow Neck, MLP has the highest mean accuracy (0.92), nearly 10%  higher than the BDT (0.83), and the results are stable. Weymouth  has the worst modelled results compared to other locations. The  best prediction in Weymouth is obtained using KNN, while the BDT  has the highest mean accuracy with relatively stable results so that  overall they could both be taken into consideration when predict-  ing at this location. Similar to Clarks, both BDT and MLP can provide  solid results for Milford with almost the same highest and mean  accuracy (0.96/0.89). However, MLP is more stable than BDT with a  smaller box size and a higher minimum accuracy value so that it is  the ﬁrst choice for predicting above and below water samples at  this location.  Our results demonstrate that there is no unique model that ﬁts  all sites and the reason can be attribute to the different conditions  and settings of each location (see also [Shaw et al., 2017](#br11)). Different  models should always be tested and the results should be  compared to look for the most appropriate solution. Model choice  also depends on which effect of the results is considered. For  example, if we are only concerned with model accuracy, KNN can  be used as it provides the best results and can predict the above and  below samples with nearly a 100% accuracy. If we care more about  the model stability, KNN then should be eliminated and BDT and  MLP-ANN should be considered because of their relatively low  variation between lowest and highest accuracy. Even SVM could be  considered as it is the most stable model for Weymouth regardless  the accuracy.  ADASYN created a balanced samples dataset by generating  ‘artiﬁcial’ above samples through a linear interpolation at the  boundary between below and above samples. With the half-faked  data set, the accuracy of model predictions signiﬁcantly  improved. However, this approach may raise some other issues  which need to be further addressed. For example, the artiﬁcial  samples could either be invading or expanding too much to the  below class, which will cause overﬁtting problem ([Luengo et al.,](#br11)  [2011](#br11)).  5. Conclusion  To help resolve the issue of unbalanced datasets in water quality  models, which result in loss of information on the above threshold  samples (exceedance) and overﬁtting for the below threshold  sample (compliance), we used an ADASYN sample balancing  method to generate artiﬁcial data of above threshold samples.  Together with machine learning techniques, this sample balancing  resulted in more accurate predictions of water quality, compared  with the original unbalanced datasets. The balanced dataset was  applied to four machine learning algorithms and resulted in much  higher accuracy in terms of both sensitivity (over 75%) and speci-  ﬁcity (over 90%) compared to past studies ([Thoe et al., 2014](#br11); [Zhang](#br11)  [et al., 2015](#br11)). K-NN, BDT, SVM, and MLP-ANN were applied to ﬁve  beach sites in Auckland and the results were compared to |  | determine which model is the most appropriate for each individual  location. Judging in terms of model accuracy, robustness and sta-  bility, BDT and MLP-ANN outperformed KNN and SVM at all  locations.  Models perform differently at different locations and many  reasons could explain this result (e.g., environmental conditions,  data quality, and model suitability). This study provided a potential  solution to the data balance issue and compared various models to  test the predictions at ﬁve sites. Future studies could focus on the  role of individual variables or on reducing uncertainties in the  synthetic generation of data. In addition, the tidal information,  which is very important for predicting FIB, is not included in this  study due to lack of detailed data. Tidal data should be added into  future models and it is likely it would further improve model  performance.  Declaration of competing interest  The authors declare that they have no known competing  ﬁnancial interests or personal relationships that could have  appeared to inﬂuence the work reported in this paper.  Acknowledgement  We want to show special thanks to the Auckland Council for  suppling of water quality data.  Appendix A. Supplementary data  Supplementary data to this article can be found online at  <https://doi.org/10.1016/j.watres.2020.115788>.  References  [Abyaneh, H.Z., 2014. Evaluation of multivariate linear regression and artiﬁcial](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref1)  [neural networks in prediction of water quality parameters. J. Environ. Health](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref1)  [Sci. Eng. 12 (1), 40](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref1).  [Anctil, F., Perrin, C., Andr](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref2)eꢀ[assian, V., 2004. Impact of the length of observed records](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref2)  [on the performance of ANN and of conceptual parsimonious rainfall-runoff](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref2)  [forecasting models. Environ. Model. Software 19 (4), 357](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref2)e[368](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref2).  [Babbar, R., Babbar, S., 2017. Predicting river water quality index using data mining](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref3)  [techniques. Environ. Earth Sci. 76 (14), 504](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref3).  [Bae, H.K., Olson, B.H., Hsu, K.L., Sorooshian, S., 2010. Classiﬁcation and regression](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref4)  [tree (CART) analysis for indicator bacterial concentration prediction for a Cali-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref4)  [fornia coastal area. Water Sci. Technol. 61 (2), 545e553](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref4).  [Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref5)  [methods for balancing machine learning training data. ACM SIGKDD Explora-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref5)  [tions Newsletter 6 (1), 20](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref5)e[29](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref5).  [Bedri, Z., Corkery, A., O’Sullivan, J.J., Deering, L.A., Demeter, K., Meijer, W.G.,](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref6)  [O’Hare, G., Masterson, B., 2016. Evaluating a microbial water quality prediction](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref6)  [model for beach management under the revised EU Bathing Water Directive.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref6)  [J. Environ. Manag. 167, 49e58](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref6).  Castelletti, A., Yajima, H., Giuliani, M., Soncini-Sessa, R., Weber, E., 2014. Planning  the optimal operation of a multioutlet water reservoir with water quality and  quantity targets. J. Water Resour. Plan. Manag. 140 (4), 496e510. [https://doi.org/](https://doi.org/10.1061/(ASCE)WR.1943&ndash%3B5452.0000348)  [10.1061/(ASCE)WR.1943](https://doi.org/10.1061/(ASCE)WR.1943&ndash%3B5452.0000348)e[5452.0000348](https://doi.org/10.1061/(ASCE)WR.1943&ndash%3B5452.0000348).  [Chan, S.N., Thoe, W., Lee, J.H.W., 2013. Real-time forecasting of Hong Kong beach](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref8)  [water quality by 3D deterministic model. Water Res. 47 (4), 1631](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref8)e[1647](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref8).  [Chandramouli, V., et al., 2007. Backﬁlling missing microbial concentrations in a](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref9)  [riverine database using artiﬁcial neural networks. Water Res. 41 (1), 217](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref9)e[227](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref9).  [Chen, W.B., Liu, W.C., 2015. Water quality modeling in reservoirs using multivariate](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref10)  [linear regression and two neural network models. Adv. Artif. Neural Syst. 6,](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref10)  [2015](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref10).  [de Brauwere, A., Ouattara, N.K., Servais, P., 2014. Modeling fecal indicator bacteria](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref11)  [concentrations in natural surface waters: a review. Crit. Rev. Environ. Sci.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref11)  [Technol. 44 (21), 2380e2453](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref11).  [Danades, A., Pratama, D., Anggraini, D., Anggriani, D., 2016. October. Comparison of](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref12)  [accuracy level K-nearest neighbor algorithm and support vector machine al-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref12)  [gorithm in classiﬁcation water quality status. In: 2016 6th International Con-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref12)  [ference on System Engineering and Technology (ICSET). IEEE, pp. 137e141](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref12).  Francy, D.S., Brady, A.M.G., Carvin, R.B., Corsi, S.R., Fuller, L.M., Harrison, J.H.,  Hayhurst, B.A., Lant, J., Nevers, M.B., Terrio, P.J., Zimmerman, T.M., 2013.  Developing and Implementing Predictive Models for Estimating Recreational  Water Quality at Great Lakes Beaches. U.S. Geological Survey Scientiﬁc |

T. Xu et al. / Water Research 177 (2020) 115788 11

|  |  |  |
| --- | --- | --- |
| Investigations, p. 68. <https://doi.org/10.3133/sir20135166/>. Report 2013-5166.  [García-Alba, J., et al., 2019. Artiﬁcial neural networks as emulators of process-based](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref15)  [models to analyse bathing water quality in estuaries. Water Res. 150, 283](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref15)e[295](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref15).  [Granata, F., Papirio, S., Esposito, G., Gargano, R., De Marinis, G., 2017. Machine](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref16)  [learning algorithms for the forecasting of wastewater quality indicators. Water](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref16)  [9 (2), 105](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref16).  [Gosain, A., Sardana, S., 2017. September. Handling class imbalance problem using](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref17)  [oversampling techniques: a review. In: 2017 International Conference on Ad-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref17)  [vances in Computing, Communications and Informatics (ICACCI). IEEE,](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref17)  [pp. 79](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref17)e[85](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref17).  [Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A., 2018. Water quality prediction using](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref18)  [machine learning methods. Water Qual. Res. J. 53 (1), 3](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref18)e[13](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref18).  [He, H., Bai, Y., Garcia, E.A., Li, S., 2008. June. ADASYN: adaptive synthetic sampling](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref19)  [approach for imbalanced learning. In: 2008 IEEE International Joint Conference](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref19)  [on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE,](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref19)  [pp. 1322](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref19)e[1328](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref19).  [He, L.M.L., He, Z.L., 2008. Water quality prediction of marine recreational beaches](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref20)  [receiving watershed baseﬂow and stormwater runoff in southern California.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref20)  [Water Res. 42 (10](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref20)e[11), 2563](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref20)e[2573. USA](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref20).  [Han, H., Wang, W.Y., Mao, B.H., 2005. August. Borderline-SMOTE: a new over-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref21)  [sampling method in imbalanced data sets learning. In: International Confer-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref21)  [ence on Intelligent Computing. Springer, Berlin, Heidelberg, pp. 878](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref21)e[887](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref21).  [Kazemi Yazdi, S., Scholz, M., 2010. Assessing stormwater detention systems treating](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref22)  [road runoff with an artiﬁcial neural network predicting fecal indicator organ-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref22)  [isms. Water Air Soil Pollut. 206 (1e4), 35e47](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref22).  [Kim, Y.H., Im, J., Ha, H.K., Choi, J.K., Ha, S., 2014. Machine learning approaches to](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref23)  [coastal water quality monitoring using GOCI satellite data. GIScience Remote](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref23)  [Sens. 51 (2), 158](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref23)e[174](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref23).  [King, R.S., Richardson, C.J., 2003. Integrating bioassessment and ecological risk](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref24)  [assessment: an approach to developing numerical water-quality criteria. En-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref24)  [viron. Manag. 31 (6), 795](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref24)e[809](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref24).  [Li, Y., Khan, M.Y.A., Jiang, Y., Tian, F., Liao, W., Fu, S., He, C., 2019. CART and PSOþ](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref25)  [KNN algorithms to estimate the impact of water level change on water quality](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref25)  [in Poyang Lake, China. Arab. J. Geosci. 12 (9), 287](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref25).  [Luengo, J., Fern](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref26)ꢀa[ndez, A., García, S., Herrera, F., 2011. Addressing data complexity for](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref26)  [imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref26)  [undersampling. Soft Computing 15 (10), 1909](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref26)e[1936](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref26).  [Marshall, M.M., Naumovitz, D., Ortega, Y., Sterling, C.R., 1997. Waterborne protozoan](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref27)  [pathogens. Clin. Microbiol. Rev. 10 (1), 67](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref27)e[85](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref27).  [Motamarri, S., Boccelli, D.L., 2012. Development of a neural-based forecasting tool](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref28)  [to classify recreational water quality using fecal indicator organisms. Water Res.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref28)  [46 (14), 4508e4520](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref28). |  | Pijanowski, B.C., et al., 2002. Using neural networks and GIS to forecast land use  changes: a land transformation model. Comput. Environ. Urban Syst. 26 (6),  553e575. <https://doi.org/10.1016/S0198-9715(01)00015-1>.  [Prakash, R., Tharun, V.P., Devi, S.R., 2018. April. A comparative study of various](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref30)  [classiﬁcation techniques to determine water quality. In: 2018 Second Interna-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref30)  [tional Conference on Inventive Communication and Computational Technolo-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref30)  [gies (ICICCT). IEEE, pp. 1501](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref30)e[1506](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref30).  [Qin, X., Gao, F., Chen, G., 2012. Wastewater quality monitoring system using sensor](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref31)  [fusion and machine learning techniques. Water Res. 46 (4), 1133](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref31)e[1144](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref31).  [Rumelhart, D., Hinton, G., Williams, R., 1986. Learning internal representations by](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref32)  [error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distrib-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref32)  [uted Processing: Explorations in the Microstructures of Cognition, vol. 1. MIT](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref32)  [Press, Cambridge, pp. 318](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref32)e[362](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref32).  [Shoaran, M., Haghi, B.A., Taghavi, M., Farivar, M., Emami-Neyestanak, A., 2018. En-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref33)  [ergy-efﬁcient classiﬁcation for resource-constrained biomedical applications.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref33)  [IEEE J. Emerg. Sel. Top. Circuits Syst. 8 (4), 693](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref33)e[707](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref33).  [Shaw, A.R., et al., 2017. Hydropower optimization using artiﬁcial neural network](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref34)  [surrogate models of a high-ﬁdelity hydrodynamics and water quality model.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref34)  [Water Resour. Res. 53 (11), 9444](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref34)e[9461](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref34).  [Stidson, R.T., Gray, C.A., McPhail, C.D., 2012. Development and use of modelling](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref35)  [techniques for real-time bathing water quality predictions. Water Environ. J. 26](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref35)  [(1), 7e18](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref35).  [Thoe, W., et al., 2014. Predicting water quality at Santa Monica Beach: evaluation of](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref37)  [ﬁve different models for public notiﬁcation of unsafe swimming conditions.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref37)  [Water Res. 67, 105](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref37)e[117](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref37).  Thoe, W., Lee, J.H.W., 2013. Daily forecasting of Hong Kong beach water quality by  multiple linear regression (MLR) models. ASCE J. Environ. Eng., 04013007  <https://doi.org/10.1061/(ASCE)EE.1943-7870.0000800>.  [Thoe, W., Wong, S.H.C., Choi, K.W., Lee, J.H.W., 2012. Daily prediction of marine](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref39)  [beach water quality in Hong Kong. J. Hydro-Environ. Res. 6 (3), 164](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref39)e[180](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref39).  [Tufail, M., et al., 2008. Artiﬁcial intelligence-based inductive models for prediction](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref40)  [and classiﬁcation of fecal coliform in surface waters. J. Environ. Eng. 134 (9),](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref40)  [789](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref40)e[799](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref40).  [Wade, T.J., Calderon, R.L., Brenner, K.P., Sams, E., Beach, M., Haugland, R., Wymer, L.,](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref41)  [Dufour, A.P., 2008. High sensitivity of children to swimming-associated](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref41)  [gastrointestinal illness: results using a rapid assay of recreational water qual-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref41)  [ity. Epidemiology 375](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref41)e[383](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref41).  [Wang, X., Zhang, J., Babovic, V., 2016. Improving real-time forecasting of water](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref42)  [quality indicators with combination of process-based models and data assim-](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref42)  [ilation technique. Ecol. Indicat. 66, 428e439](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref42).  [Zhang, Z., et al., 2015. Modeling fecal coliform bacteria levels at Gulf coast beaches.](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref43)  [Water Qual. Expo. Health 7 (3), 255](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref43)e[263](http://refhub.elsevier.com/S0043-1354(20)30325-0/sref43). |

**译文：**

|  |  |  |
| --- | --- | --- |
| 一种基于自适应合成采样算法和机器学习的休闲水质预测模型 徐汀汀\* ，Giovanni Coco, Martin Neale  新西兰奥克兰大学理学院环境学院 | |  |
| 文章信息 | 摘要  预测休闲水质是水资源管理中最困难的任务之一，对人类和社会具有重大影响。许多数据驱动的模型已被用于预测水质 指标， 以便能够实时评估公共卫生风险。这种评估最常基于粪便指标细菌（FIB） ，将FIB的值与指南中公布的阈值相比 较。然而，在水质数据集中，FIB值往往不平衡，超过指导阈值的数据比例很小，而数量更大。这可能是接受模型预测 的一个限制因素，因为即使总体精度很高，预测的敏感性也可能很低。为了解决这一问题，本文提出了一种自适应合成 采样算法（ADASYN）来生成合成的阈值以上FIB实例，并验证了该方法预测休闲水质的有效性。本文的模型基于四种机 器学习技术：k-均值最近邻、增强决策树、支持向量机和多层感知器人工神经网络，并应用于新西兰奥克兰的五个不同 地点。除了支持向量机外，所有模型都提供了相对较高的灵敏度（约75%）和总体精度（超过90%） ，这表明通过使用涉 及人工数据的更复杂的模型训练，可以有效地预测合规条件和超越条件。考虑到模型的精度和稳定性，增强决策树（BD T）和多层感知器人工神经（MLP-ANN） 网络是最好的两个模型，多层感知器效率最高，计算时间最短。  ©2020爱思唯尔有限公司。保留所有权利。 | |
| 文章历史记录：  收到2019年12月9日 以修订的形式收到  2020年4月1日  2020年4月2日接受  2020年4月13日在线提供 |
| 关键词： 水质  自适应合成采样算法  最近的邻居  启动决策树  支持向量机  人工神经网络 |

**1.介绍**

关于海滩水质的可靠和可靠的信息至关重要，因为它被用来向管理人 员了解水中娱乐带来的公共健康风险（Thoe和Lee，2013；Stidson等人， 2012；Francy等人，2013；韦德等人，2008年）。海滩水质通过粪便指示 菌广泛监测（FIB）进行监测，这通常涉及测量水中肠球菌或大肠杆菌的 浓度。监测结果通常被评估是否符合指南，并且指南超出会触发海滩使用 者的警告或警告（deBrauware等人，2014；卡斯特莱蒂等人，2014；马歇 尔等人，1997年；金和理查森，2003年）。FIB的实验室检测通常需要24 小时，使用最可能的方法（MPN），然而，一旦样本收集和包括对结果的解释和交流，测试结果可能在样本收集后的48小时内不公开 。这导致水质信息可能过时，因为水质的变化可能比分析时间更快。出于 这个原因，人们投入了大量的精力来实时预测FIB浓度，利用一系列基于 过程和数据驱动的模型(e。g.Feng等人，2010年；Abyaneh，2014年；Thoe 和Lee，2013年；He和He，2008年)。用于预测水质的绝大多数模型是数据驱动的、统计和分类的，因此对 开发中使用的数据敏感（Thoe等人，2014；Chen和Liu，2015；Li等人，2019）。例如，所有这些模型都面临着数据集不平衡的问题，即符合指导 方针的数据点的数量总是远远超过超过FIB的指导方针阈值的数据。因此 , 该模型通常善于预测顺应条件，而不是超越条件。钱德拉莫利等人。 ( 20 0 7 ） 和Tu fail al 等 人。

\*通讯作者。

电子邮件地址：txu648@奥克兰德大学。ac.nz (T.徐）。

[https://doi.org/10.1016/j.瓦特雷斯.2020.115788](https://doi.org/10.1016/j.watres.2020.115788)

0043-1354/ ©2020 爱思唯尔有限公司 。保留所有权利。

2 T.徐等人。/水研究177（2020） 115788

（2008）使用人工神经网络（ANN）来预测水质，发现它们略优于传统的 回归模型。然而，他们只考虑了总体精度，这可能导致由于原始数据集包 含了不成比例的兼容数据点(i。e., 当FIB低于指导方针的阈值时)。2010 年，Kazemi Yazdi和Scholz通过神经网络模型基于地表径流评估水质，并 将神经网络与多线性回归模型（MLR）进行比较。人工神经网络在预测径流处理方面优于MLR，因为它有更高的能力捕捉水 质和微生物因素之间的非线性关系。他们专注于探索质量指标和影响因子 之间的关系，但没有提供模型验证方面的细节。秦等人。（2012）引入了 一种先进的基于增强树的机器学习模型，但没有将它们的结果与其他模型 进行比较。Thoe等。（2014）采用了三种回归模型以及分类树（CT）和神 经网络。他们评估了模型结果的总体准确性，但也评估了敏感性和特异性 , 并得出结论，CT和ANN的表现优于回归模型。这里的总体准确性是指模 型成功预测阈值以下和阈值以上的数据，敏感性评估模型仅预测超过数据 的能力，特异性评估符合性数据的预测。即使在这种情况下，敏感度仍然 很低，可能是由于对训练数据集中更不超过的数据的偏差（Bedri等人，2016；Wang等人，2016；Shaw等人，2017）。除了这些模型，多层感知器 神经网络预测水质的能力(Zhang等，2015；Garcia-Alba等，2019年）。然而，这些研究认为人工神经网络是一个可以集成到其他模型中的组成部 分，并没有与其他机器学习技术进行比较。格拉纳塔等人。（2017）和Haghiabi等人。（2018）使用了几种机器学习方法来预测水质，并报告了 对良好水质的显著高估，但没有探索推理。此外，他们用于训练网络的数 据集仍然不平衡。

众所周知，人工神经网络等机器学习方法需要一个高质量的训练数据 集来提供可靠和准确的结果。2012年，莫塔马里和博切利开发了一种机器 学习模型，用来对河流尺度上的休闲水质进行分类。他们发现，基于神经 的学习向量量化模型的表现优于MLR和ANN，并发现所有的预测都严重依赖 于训练样本特征的有效性。然而，休闲水质数据集的最大问题仍然不存在 , 即它们总是不平衡，绝大多数数据集低于指导阈值，只有非常少的数据 超过指导阈值（Thoo等，2012；Garcia-Alba等，2019年）。当训练数据 驱动的模型时，一个不平衡的数据集增加了少数类的信息丢失和多数类的 过拟合的可能性，这不能仅仅通过减少或重复数据样本来轻易解决（Bati sta et al.，2004）。金等人。（2014）也指出了在尝试使用三种机器学 习方法使用卫星数据监测水质时存在的这个问题。有限的训练数据，特别 是对恶劣的水条件，削弱了预测。为了解决不平衡的样品问题，合成少数 过采样技术（烟雾）常用于其他领域（Han等，2005；Luengo等，2011）。SMOTE通过在多数阶级和少数阶级之间进行插值来生成人工样本。然而,使用这种方法，“合成 ”样本仍然更有可能分布在少数类的内部（Luen go et al.，2011）。这一限制可以通过先进的自适应合成采样算法（ADASYN）来克服。ADASYN , 可以在这两个类之间的边界上创建更多的样本，提高训练精度（He et al.，2008；Gosain和Sardana，2017）。然而，在已发表的ADASYN研究中，数 据维度通常较低，不超过三个输入变量。因此，它在高维数据上的效率仍 有待评估。此外，大多数研究集中在一个地点和一个模型，或最多在一个 地点比较两种类型的模型，通常比较人工神经网络和线性回归（Thoe等人 , 2012；Chan et al.，2013）。Thoe等。（2014）和Danades等人。（20 16）通过比较三种或更多不同的模型来揭示这个问题，以揭示它们的优缺 点，但只在一个位置。其他研究集中于多个地点，但只考虑了一个模型 ( Garcia-Alba等人，2019年）。因此，尚未对使用不同位置和不同算法进 行灵敏度分析。

为了克服上述缺陷，我们的研究提供了四种机器学习算法预测FIB在新 西兰奥克兰五个海滩的预测的全面比较。 目的是： (1)应用ADASYN改进预 测水质数据集的训练，解决水质数据集不平衡的问题； (2)在使用平衡数 据集时评估机器学习模型的能力； (3)使用不同的机器学习模型比较不同 位置的不同模型结果。通过完成这些目标，我们试图回答两个问题：我们 是否可以使用人工水样来平衡机器学习模型的原始数据集？那么，一个平 衡的数据集是否能提高预测能力呢？

**2 . 研究区域和数据** 2.1.研究地点

这五个研究地点，米尔福德海滩、窄颈海滩、法官湾、韦茅斯海滩和 克拉克海滩，分布在新西兰奥克兰的海岸线上，具有独特的地理条件(图1 ) . 前两个是开放的海岸海滩，面向太平洋，以海浪为主。其余的三 个地点是在隐蔽的海湾和港口，所以主要是潮汐过程。2017年，奥克兰议 会启动了一项基于修订模式的海滩水质计划（ “安全用水 ”）。模型的使 用是基于它们能够满足美国地质调查局发布的性能标准的能力（USGS eFrancy et al.，2013）。基于此指导，准确率超过85%、灵敏度为50%、特异性为80%的模型应被认为是在标准水平上预测水质的合格模型。为了 将水质预测模型纳入安全管理平台，它必须在总体准确性、灵敏度（对高 于阈值的预测能力）和特异性（对低于阈值的预测能力）方面满足基准。 模型处于不断的审查状态，并定期改进，以响应奥克兰议会正在进行的海 滩采样。

2.2.数据

5个研究地点的1995年至2018年的长期水质监测数据，以及伴随的环境 变量（2019年也适用于Clarks）。每个数据集包含8个变量。2): FIB，用 作确定水质阈值（单样本阈值¼280、FIB司280：超越和FIB < 280：合规 ) , 新西兰指南规定，二十四、四十八、七十二小时（一、二、三天）， 累计降水量、风速和每天太阳小时。图2显示韦茅斯有

T.徐等人。/水研究177（2020） 115788 3

|  |
| --- |
| 地图  描述已自动生成 |

图1. 新西兰奥克兰的五个水监测点（每张图像中的黑点表示采样位置）。

与其他四个位置相比，最平衡的数据集，而窄颈有最不平衡的数据集，有 少于2.5%的样本被发现高于阈值。克拉克和韦茅斯的风向分布均匀，但其 他三个地方经历了更多的南风。米尔福德的平均太阳小时数是最小的，大 约每天6.5小时，法官湾的平均日照时间最长，几乎每天7小时。韦茅斯地 区72 h内的平均降雨量最大，接近8.2 cm，而窄颈地区的平均降雨量最小 , 72 h内的平均降雨量为5.6cm(见图。3).

**3 . 方法学**

与回归模型通常预测绝对FIB值作为连续函数，然后根据阈值重新分类 该值不同，四种机器学习算法通过二元分类过程来“预测 ”水样的FIB是 否超过阈值。

3.1 K-最近邻算法（K-NN）

k-最近邻算法是一种用于分类和回归的非参数方法。在这两种情况下,输入由特征空间中的k个最接近的训练示例组成。邻居取自一组对象，其中的 类（kNN分类）或对象属性值（k-NN回归）是已知的。3 ) . 这种方法被广 泛应用于分类问题（Babbar和Babbar，2017）。

3.2.提升决策树（BDT）

决策树是一种类似于流程图的结构，其中每个内部节点表示对一个属 性的“测试 ”(e。g.无论抛硬币是头还是尾)，每个分支代表测试的结果 , 每个叶节点代表一个类标签（在计算所有属性后做出的决定）。从根到 叶的路径表示分类规则。增强法是一种将许多弱学习者（树）组合成一个 强分类器和增强决策树的方法。4)因其出色的准确性和快速操作而受欢迎 （Prakash等，2018；Sshhinan等人，2018）。

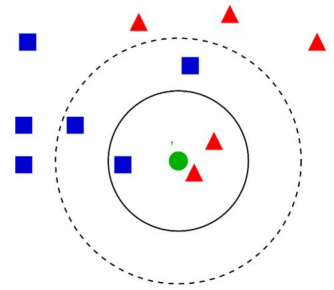
.3.3.支持向量机（SVM）

支持向量机（SVM）是一种由一个分离的超平面正式定义的判别分类器 。换句话说，给定标记的训练数据，该算法输出一个最优超平面

图表, 雷达图

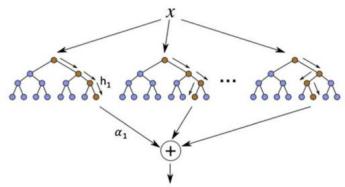
描述已自动生成

图2.在奥克兰的5个地点设置的数据集。使用极坐标系，方向表示风向（0e360o） ，与原点的距离表示每日太阳小时（0e15 h） ，点的大小反映72 h内（0e100厘米）累积的总降水量，颜色 表示水质是否超过阈值（蓝色：FIB司280和绿色：FIB < 280） 。 (为了解释本图例中对颜色的引用，读者请参考本文的Web版本。)



|  |
| --- |
|  |

图3.k-NN分类示例。测试样本（绿点）应分为蓝色正方形或红色三角形。如果k¼3（实线圆 ) , 它被分配给红色三角形总体，因为内圆内有2个三角形，只有1个正方形。如果k¼5（虚 线圆） ，它被分配给蓝色的方形人口(3个方形vs。外圆内有2个三角形)。（要解释本图例 中提到的颜色，请参考本文的Web版本。）



|  |
| --- |
|  |

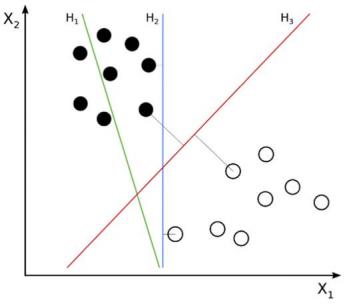
图4 . 提升决策树：h是弱分类器，a表示分配给h的范围权值。

它分类了新的例子。在二维空间中，这个超平面是一条线，将一个平面分 成两部分，每个类都位于不同的一边。5).

3.4.人工神经网络（ANN）

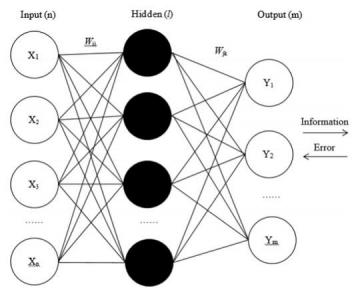
人工神经网络是一种广泛使用的建模技术，具有自适应、 自组织和自 学习能力（Pijanowski等，2002年，Anctil等，2004年）。人工神经网络 包括一个输入层（变量输入算法）、一个或多个隐藏层（输入变量组合） 和一个输出层（预测）。由于其简单、易于训练以及合理的联想记忆和预 测能力（Rumelhart et al.，1986），我们使用了前馈、错误反向传播三 层感知器（BP-TLP）ANN架构(图。6).BP-TLP最重要的特点是，在训练阶 段，信息是向前传递的，而错误是反向传播的。在前馈转换过程中，输入 层的信息（值/特征）由隐藏层进行处理，然后转换回输出层。在一个图 层中的每个节点

T.徐等人。/水研究177（2020） 115788 5



|  |
| --- |
|  |

图5.SVM分类的一个例子。H1并不分离这些类。h2可以，但只有很小的边际。H3以最大的边 际分隔它们。



|  |
| --- |
|  |

图6.BP-TLP ANN的体系结构

这样将会影响下一层中的节点状态。如果输出层的值不满足预期的结果， 则会转向反向传播过程。根据预期误差，BP-TLP神经网络开始调整网络的 权值和阈值，允许预测结果接近预期结果。

3.5.自适应合成采样算法（ADASYN）

ADASYN是合成少数实例过采样技术（SMOTE）的改进版本，该技术用于 避免在将少数实例的精确副本添加到主数据集时发生过拟合（Gosain和

Sardana，2017）。ADASYN算法的关键思想是使用密度分布作为标准， 自 动确定为每个少数数据示例需要生成的合成样本的适当数量。

6 T.徐等人。/水研究177（2020） 115788



|  |
| --- |
|  |

图7.在打击的帮助下生成合成实例。a-e为合成样品，x1-x6为监测样品。

密度分布可以由基于n维向量的多数样本和少数样本间的欧氏距离的knn函 数得到。从少数类中提取数据子集作为例子，然后创建新的合成类似实例 。7).SMOTE在这里停止，但是使用ADASYN，一个随机的小值被添加到“伪造的”样本中，使新样本更真实，而不是与原始样本线性相关（关于ADASYN应用程序的更多细节可以在附录I中找到）。

3.6 建模框架

图8说明了我们使用的工作流，展示了数据是如何被处理和应用到模型 中的。采样数据构成原始数据库，包括两部分为七个训练字段和一个标签 字段（FIB）。使用FIB值，将样本分为两类（FIB高于或低于280），以保 持与新西兰指南的一致性。然后应用ADASYN生成高于阈值的合成样本，以 平衡数据集。为了形成训练数据集，从阈值为280个以下的样本中随机选 择数据样本，并与合成的样本相结合。然后将新的训练样本数据集作为输 入数据作为输入数据输入到四种机器学习算法中。将阈值以下的未选择的 剩余样本（<280）与原始样本（不是通过ADASYN生成的）阈值以上样本 ( >280）耦合，形成验证和测试数据集，并应用于模型结果的验证过程。

利用Matlab实现了k-最近邻（KNN）、增强决策树（BDT）、支持向量 机（SVM）和多层感知器（MLP-ANN）。前三个方法采用优化的超参数和贝 叶斯优化过程来寻找最佳参数。贝叶斯优化是Matlab的一个内置函数，可 以自动优化最关键的参数（超参数）。优化过程可以通过改变



|  |
| --- |
|  |

图8. 建模框架的流程图。在括号中是样本的数量。请注意：n¼naþ nb ; nd¼ nb- na.

参数迭代。对于每个单独的机器学习算法，关键参数是不同的。表1描述 了关键参数和优化结果。而MLP-ANN则采用不同的方法来决定参数。首先，隐藏节点的数量会显 著影响人工神经网络的性能。节点过少会导致显著的预测误差，而过多的 节点会延长训练过程，导致过拟合。基于规则，隐藏的神经元的数量不应 该小于2n/3þ1（n¼输入节点的数量），我们尝试数字6、7、8、9，和隐藏 的神经元的最佳数量，8，是我们最后的选择的模型性能和网络的简单性 。然后，在Matlab中实现了莱文伯格马夸特算法函数来训练网络更新权重 和偏差，因为其强大的计算相对较小的数据集和高效的反向传播。此外， 隐层和输出层的传递函数分别为s型（tansig）和线性函数（in）。每个模型对每个研究地点运行100次，以评估模型的能力和稳定性。使 用混淆矩阵来计算每个模型的总体准确性、灵敏度和特异性。同时还记录 了运行时间来评估计算效率。

总体上，如图所示。8，对于有n个样本的数据D，na和nb是高于和低于 FIB阈值的水样数量，其中naplusn等于n。注意，na比nb小得多，它们的 区别是nd。在我们的研究中，合成样品的总需求数也等于nd。然后，我们 从nb中随机选择阈值以下的样本，并将其与

T.徐等人。/水研究177（2020） 115788 7

表1 超参数可以通过贝叶斯优化的方法进行优化。

|  |  |  |  |
| --- | --- | --- | --- |
| 方法名称 | 超参数 | 描述 | 结果(e。g.战斗开始 与防御装备 |
| 克恩 | 距离度量 | 最近邻的距离搜索方法 | 马哈拉诺比斯 |
| 邻居 | 用于对每个点进行分类的最近邻居的数量 | 8 |
| BDT | 提高方法学习周期数学习率 | 集成聚合方法，特别关注不同的增强函数学习的学习总数 | 温和的提升 |
| 最小叶片尺寸 | 学习过程的收缩范围  叶节点观测值的最小数量 | 43  0.0099 |
| SVM | 框约束 核尺度 | 对违反保证金的观察结果所施加的最高惩罚 一个用于划分预测器矩阵的所有元素的值 | 81  0.0089 0.001 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 等量的合成样本制作新的训练水样数据集（ntrain¼2）。原始的阈值以上 样本与剩余的以下样本（等于nb - nd）相结合，用于模型验证。在将建 模框架应用于五个地点之前，还采用了简化的下和以上采样方法，以减少 以下样本的数量，并等于以下和以上样本的数量。  **4 . 结果与讨论**  4.1.原始数据集  表2描述了水质监测数据样本，表3显示了使用原始数据集的所有算法 的训练结果。表2显示，所有监测点的水样都很不平衡，只有韦茅斯在阈值上下的样本数量上平衡得稍好一些。从表3中可以看出，尽管总体精度 较高，但由于上述阈值样本（超过阈值）均未被成功预测，因此四种ML方 法对这5个位置都不能很好地建模。由于数据集的不平衡问题，这种高模 型的总体精度可能会有问题。由于所有的数据集出现在阈值以上的样本非 常有限，算法只关注阈值以下的样本。每个模型的总体精度都较高，这是 由于对阈值以下样本的正确预测（符合性），而上述样本则没有得到很好 的预测。因此，这四种模型都可以精确地预测以下样本，但没有一种算法 能够识别FIB超过280阈值的水样。  4.2.过采样方法  在应用ADASYN之前，首先使用下采样方法来平衡数据。我们通过一个 随机选择的过程中减少了以下样本的数量，直到它等于上述样本的数量。 在此基础上，在采样数据集下的结果（表4）表明，上述四种算法都可以 更准确地预测水质。KNN和BDT对所有四个位置都有相同的结果，并且预测 来自Clarks的样本组的总体精度为0.75。SVM和ANN也有助于预测来自韦茅 斯和米尔福德的水样组，总体精度分别为0.73和0.75。这些结果比使用原 始数据更有意义  表2 来自五个监测点的水质样本（高于/低于阈值）。 | | | 表3 使用原始数据集进行预测。 | | | |
| 现场 机器学习    测试符合。 以上 在 … | | | |
| 克拉克斯 窄颈  法官湾  韦茅斯  米尔福德 | 0.93 0.94 0.94 0.87 0.99 | 0/5 0/8 0/8 0/11 0/2 | 68/68  135/135 132/132 73/73  162/162 |
| 因为上面和下面的样本都可以被预测出来。然而，其他模型明显低估了不 同地点的以下样本。以米尔福德为例，SVM低估了上面的三个样本，而高 估了下面的两个样本。  对以下样本的低估是由于欠采样方法丢失了潜在有用的信息。训练样 本可能不够充分，而且它可能不是低于阈值的样本的一个代表性例子。因 此，即使整体精度是可以接受的，结果也是不准确的，不足以用于预测水 质。过采样方法与欠采样方法相反。对于每个位置，我们复制上面的样本 , 然后将它们与下面的样本组合起来，生成一个新的数据集。过采样预测 的结果见表5。  在表5中，我们可以看到BDT在总体精度方面优于其他三种算法，而SVM 表现最差。除韦茅斯外，BDT、KNN和ANN在所有位置都显示出相对较高的 建模精度。然而，结果仍然具有欺骗性，因为只有复制了相同的上述样本 与不变的样本值可能会增加过拟合的可能性。此外，如果在阈值以上的样 本中存在一些误差，过采样方法只会放大偏差。  4.3.ADASYN方法  采用自适应合成采样算法（ADASYN）来改善数据平衡。ADASYN通过线 性插值从上述样本中综合创建新的样本。这种方法在两个类之间的边界附 近比在上述样本的内部创建了更多的样本。这种方法 | | | |
| 阈值¼280 | 克拉克斯 | 窄颈 | 贝韦茅斯法官 | 米尔福德 | | |
| 司280 | 14 | 15 | 23 | 74 | | 28 |
| <280 | 231 | 463 | 444 | 207 | | 520 |
| 合计 | 245 | 478 | 467 | 281 | | 548 |
| 比率（上图） | 0.0571 | 0.0314 | 0.0493 | 0.2633 | | 0.0511 |

表4 采样不足的结果。

现场

KNN和BDT

SVM

安



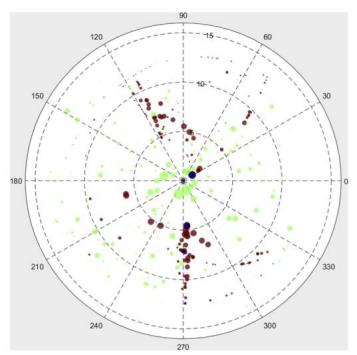
Acc。下面是Acc..的上方。下面是Acc..的上方。下图

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 克拉克斯 | 0.75 | 3/3 | 3/5 | 0.50 | 3/3 | 1/5 | 0.38 | 3/3 | 0/5 |
| 窄颈 | 0.56 | 3/3 | 2/5 | 0.33 | 3/3 | 0/6 | 0.33 | 3/3 | 0/6 |
| 法官湾 | 0.50 | 6/8 | 1/6 | 0.57 | 8/8 | 0/6 | 0.50 | 6/8 | 1/6 |
| 韦茅斯 | 0.52 | 11/12 | 12/32 | 0.73 | 0/12 | 32/32 | 0.43 | 9/12 | 10/32 |
| 米尔福德 | 0.44 | 6/7 | 1/9 | 0.69 | 4/7 | 7/9 | 0.75 | 7/7 | 5/9 |

缓解了人工生成合成新样本时产生的重复过采样引起的过拟合问题，从而 不造成有用信息的损失，减少过拟合的风险。图9说明了用ADASYN方法获 得的数据集。

我们首先使用ADASYN数据集应用了这四种ML算法，并使用Clarks湾的 原始数据集进行了验证和测试。结果明显改善，见图。10（我们以MLP ANN为例展示了结果，其他算法的结果可以在附录II中找到）。在无花果 。10（1、1）和（0、0）表示有多少水样被正确预测/分类。第一个值表 示对高于阈值的水样的正确预测（灵敏度），第二个值反映了以下样本建 模的准确性（特异性）。，高估和低估用（0、1）和（1、0）表示。KNN 使用ADASYN数据集进行训练，并使用原始数据集进行验证，精确预测上述 阈值和以下阈值样本。总体准确率为88.43%。特异性（低于样本）和灵敏 度（高于样本）的准确率分别为89.25%和76.47%，均高于样本下方法和其 他模型。在这个平衡的数据集下，低估和高估也是可以接受的，因为只有 10.75%的以上阈值样本被遗漏，23.53%的以下样本被认为是错误的。与其 他研究相比，特异性准确率没有显著提高，但敏感性预测准确率提高到70%以上。过采样和ADASYN之间的区别在于数据样本本身。对于过度抽样，我们 只需重复现有的少数类样本（高于280），直到这个数字等于大多数类 ( 低于280）。例如，假设有200个样本低于阈值，280个，只有20个样本的 值大于280。简化的过采样将只复制这20个以上的样本10次，直到达到与 以下样本相同的数量，这些重复的样本将输入训练和验证过程。因此，由 于在训练和验证过程中使用相同的数据样本，最终的模型验证结果可能会 很高。此外，这种方法还放大了过拟合。

然而，使用ADASYN方法，将通过产生新的值，而不是简单地复制现有 的值，将数据集与全新的人工阈值以上数据样本进行平衡。训练数据集和 验证数据集之间没有冲突。合成的样本将只用于训练模型，而真实的数据将只应用于模型验证基于以上解释，即使过采样的精度值高于ADASYN，我们也不能说过采 样模型的性能优于ADASYN。相比之下，来自ADASYN的结果更可靠，因为数 据集是平衡的，并通过创建人工数据来分离训练和验证数据集。



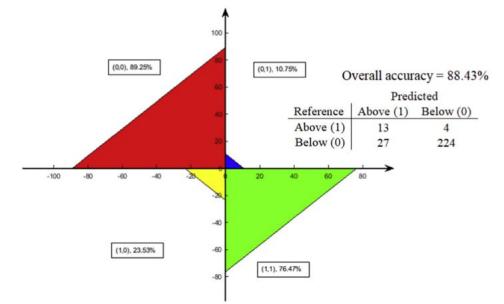
|  |
| --- |
|  |

图9.Clarks海滩的平衡数据，绿点：下面样本，蓝点：上面样本，棕色点： ADASYN样本。 方向、距离中心的距离和点的大小代表风向（0e360o）、每日太阳小时（0e15 h） ， 以及 72 h内累积的总降水量（0e100 cm） 。（要解释本图例中提到的颜色，请参考本文的Web版 本。）

表5 过采样结果。

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 现场 | 克恩 | | BDT | | | | SVM | 安 | | | |  |
|  | Acc。 | 以上 | 在…下面 | Acc。 | 以上 | 在…下 面 | Acc。 | 以上 | 在…下面 | Acc。 | 以上 | 在…下面 |
| 克拉克斯 | 0.98 | 72/72 | 66/69 | 0.99 | 64/64 | 76/77 | 0.69 | 46/68 | 51/73 | 0.94 | 59/59 | 43/49 |
| 窄颈 | 0.98 | 147/147 | 126/132 | 0.99 | 142/142 | 136/137 | 0.68 | 105/142 | 85/137 | 0.97 | 102/102 | 105/112 |
| 法官湾 | 0.96 | 127/127 | 128/138 | 0.99 | 132/132 | 131/133 | 0.67 | 64/131 | 114/134 | 0.91 | 80/80 | 104/123 |
| 韦茅斯 | 0.83 | 65/68 | 42/61 | 0.88 | 62/68 | 51/61 | 0.64 | 47/73 | 36/56 | 0.73 | 31/47 | 41/52 |
| 米尔福德 | 0.98 | 154/154 | 151/157 | 0.99 | 163/163 | 145/148 | 0.73 | 93/151 | 135/160 | 0.97 | 109/109 | 123/130 |

T. Xu等人。/水研究177（2020） 115788 9



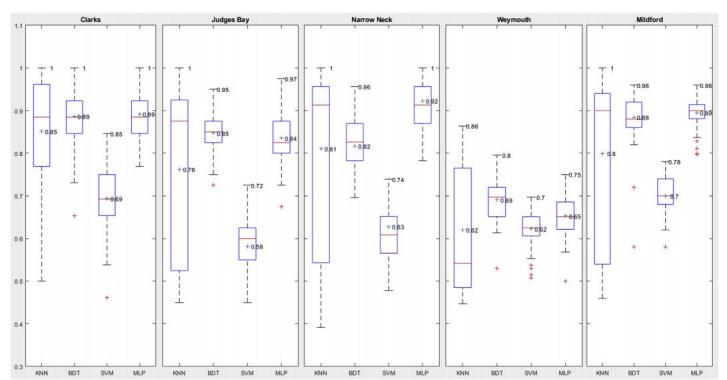
|  |
| --- |
|  |

图10. 使用ADASYN，Clarks海滩可视化MLP-ANN结果。其他算法的计算结果见附录II。

4.4 比较了5个地点之间的4个模型结果

图11（在本文的结尾）使用具有ADASYN数据集的四种ML算法，显示了 所有位置的模型和原始数据之间的验证结果。箱线图，用于显示模型的精 度（我们对每个算法执行了100次运行），显示了最高的精度（底线）、平均精度（中线）、最低的精度（底线）和精度范围（盒大小）。红叉表 示模型预测中的任何最终异常值。KNN精确地预测了所有样本的上下样本

海滩的平均准确率在80%左右。除韦茅斯外，至少有一个KNN可以正确预测 上下样本，接近达到100%的准确率。然而，KNN的盒子大小总是大于其他 三种模型（每个位点的最低精度总是与KNN相关），这表明KNN的100次运 行存在很大的变化。我们得出的结论是，这些结果不像其他模型那样稳健 和可靠。SVM是性能最差的算法，平均模型精度值很低，从不高于0.7。这 个模型不应该被认为是一个合适的预测模型，因为它在所有位置的性能都 较差(一些基于SVM



|  |
| --- |
|  |

图11.箱线图显示了建模精度的五个监测点。

10 T.徐等人。/水研究177（2020） 115788

预测低于0.5)。BDT和MLP-ANN预测比KNN和SVM(图中相对较小的盒子尺寸 。1 1).BDT和MLP-ANN的结果与克拉克、韦茅斯和米尔福德的结果非常相 似。BDT在法官湾上略优于人工神经网络（盒子尺寸更小，平均值更高，为0.85，最小精度值为0.70），然而，人工神经网络在窄颈上提供了更好 的结果，平均精度值（0.92）远高于BDT（0.82）。在不同的位置上，模型的性能也有所不同。BDT和MLP都可以用于预测Clarks海滩的上下水样（平均精度高，0.89，稳定性好，小盒尺寸）。对 于法官Bay，BDT应该被认为是最合适的模型，因为平均精度值第二高（0.85），所有模型之间的变化最小。对于窄颈，MLP具有最高的平均精度（0. 92），比BDT（0.83）高出近10%，结果是稳定的。与其他地方相比，韦茅 斯的模型结果最差。韦茅斯的最佳预测是使用KNN获得的，而BDT的平均精 度最高，结果相对稳定，因此在该位置进行预测时，总体上可以考虑到两 者。与Clarks类似 ，BDTAN和MLP都可以为米尔福德提供可靠的结果,具有几乎相同的最高和平均精度（ 0.96/0.89 ） 。然而，MLP比BDT 更稳定，盒子尺寸更小，最小精度值更高，因此它是预测该位置上下水样的首选。

我们的研究结果表明，没有一个独特的模型适合所有的地点，其原因 可以归因于每个地点的不同条件和设置（另见Shaw等人，2017年）。应该 始终测试不同的模型，并应该比较结果，以寻找最合适的解决方案。模型 的选择也取决于所考虑的结果的影响。例如，如果我们只关注模型的准确 性，那么可以使用KNN，因为它提供了最好的结果，可以预测上下样本，准确率接近100%。如果我们更关心模型的稳定性，那么就应该消除KNN，并且应该考虑BDT和MLP-ANN，因为它们在最低和最高精度之间的变化相对 较低。即使是SVM也可以被认为是韦茅斯最稳定的模型，无论精度如何。ADASYN通过在下面和上面的样本之间的边界上进行线性插值来生成“ 人工的 ”上面的样本，从而创建了一个平衡的样本数据集。在半伪造的数 据集下，模型预测的准确性显著提高。然而，这种方法可能会提出一些需 要进一步解决的其他问题。例如，人工样本可能会入侵或扩展到下面的类 别，这将导致过拟合问题（Luengo et al.，2011）。

**5 . 结论**

为了帮助解决水质模型中数据集不平衡的问题，导致上述阈值样本信 息丢失（超过）和阈值以下样本过拟合（合规），我们使用ADASYN样本平 衡方法生成阈值以上样本的人工数据。与机器学习技术一起，与原始的不 平衡数据集相比，这种样本平衡导致了更准确的水质预测。平衡数据集应 用于四种机器学习算法，与过去的研究相比，在灵敏度（超过75%）和特 异性（超过90%）方面都获得了更高的准确性（Thoo等人，2014；Zhang等 人，2015）。将K-NN、BDT、SVM和MLP-ANN应用于奥克兰的5个海滩地点， 并将结果进行了比较，确定哪个模型最适合于每个单独的位置。从模型精度、鲁棒性和稳定性来 看，BDT和MLP-ANN在所有位置都优于KNN和SVM。模型在不同位置的表现不同，许多原因可以解释这一结果(e。g., 环 境条件、数据质量和模型适用性)。本研究为数据平衡问题提供了一个潜 在的解决方案，并比较了不同的模型来测试五个站点的预测。未来的研究 可以集中于单个变量的作用，或减少数据综合生成中的不确定性。此外， 潮汐信息对预测FIB非常重要，由于缺乏详细的数据，本研究未纳入本研 究。潮汐数据应该被添加到未来的模型中，它很可能会进一步提高模型的性能。