

Predicting Shellfish Farm Closures with Class Balancing Methods

Claire D'Este¹, Ashfaque Rahman¹, and Alison Turnbull²

¹ Intelligent Sensing and Systems Laboratory and Food Future Flagship, CSIRO,
Castray Esplanade, Hobart, Australia 7000

² Department of Health and Human Services,
GPO Box 125, Hobart, Australia 7001
`claire.deste@csiro.au`

Abstract. Real-time environmental monitoring can provide vital situational awareness for effective management of natural resources. Effective operation of Shellfish farms depends on environmental conditions. In this paper we propose a supervised learning approach to predict the farm closures. This is a binary classification problem where farm closure is a function of environmental variables. A problem with this classification approach is that farm closure events occur with small frequency leading to class imbalance problem. Straightforward learning techniques tend to favour the majority class; in this case continually predicting no event. We present a new ensemble class balancing algorithm based on random undersampling to resolve this problem. Experimental results show that the class balancing ensemble performs better than individual and other state of art ensemble classifiers. We have also obtained an understanding of the importance of relevant environmental variables for shellfish farm closure. We have utilized feature ranking algorithms in this regard.

1 Introduction

Authorities such as the Tasmanian Shellfish Quality Assurance Program (TSQAP) are responsible for ensuring that shellfish harvested from commercial growing areas are shown to be free of harmful contaminants. Microbial contaminants in particular pose a major risk to public health. The potential presence of such contaminants are continually monitored through a combination of manual sampling and real-time sensors at each of the shellfish growing sites through indicators of salinity, rainfall and river flow, as well as direct water samples of the microbes themselves. When these indicators fall outside the limits provided by the management plan at a particular site, the growing site must be closed. If the shellfish are close to harvesting size and the closure time is lengthy this can result in significant loss of stock. This has economic implications for the individual farms, but can also cause disruptions in the supply of shellfish causing negative impact on the industry statewide.

Management of the TSQAP is hindered by tedious and complex manual processes that are required to obtain the relevant environmental data. This is mainly

due to different sensors residing with different organisations that provide the data in different formats through different means; including FTP servers, dial-up communications, web pages etc. This has resulted in a system that relies heavily on the availability of a single expert and has resulted in harmful events being missed when this expert was unavailable. Consequently, we have begun developing a real-time decision support tool for the TSQAP. This tool will deliver all of the proxy indicators of quality assurance from each of the different data custodians in real-time and combine them within a visual interface to provides an overall picture of the water quality at each site. In addition to the initial improvements obtained through data integration, the use of data mining and knowledge discovery may potentially assist with short-term and long-term management decisions.

In this paper we present a classification approach to predict shellfish farm closures. Shellfish farms are closed depending on location, rainfall, salinity and toxicity etc. We have developed a data set where we obtained information on farm closures and corresponding readings of environmental variables. This gives rise to a classification problem. However as the closures are relatively infrequent events, traditional techniques will tend to favour the majority class resulting in a system that always predicts that no event is present. As well as the issues we will face in the future with regards to unreliable inputs from sensors and heterogenous sensors at each of the farm locations, we first need to approach the minority class problem. This will also be relevant when we analyse other potentially harmful events to the aquaculture industry such as algal blooms. We have developed a random class balancing ensemble method to address this problem. We have also presented the results of feature ranking to understand of the relative importance of relevant environmental variables for shellfish farm closure.

2 Related Work

Shellfish farm closure prediction remains to date a novel application of data-driven techniques. In general, data mining/machine learning techniques are rarely applied to aquaculture problems; with the exception of prediction of harmful algal blooms [1].

Ensemble classifiers [2] create complementary base classifiers by manipulating the training set. The idea is to train base classifiers on different subsets of the data. In bagging [3–9] the subsets are randomly drawn (with replacement) from the training set. Boosting [10–13] is a hierarchical process where the first subset is created by randomly drawing patterns from the training set. The patterns that are not correctly classified by the current classifiers are given more importance by the classifiers following passes.

Synthetic Minority Oversampling Technique (SMOTE) boosting was developed to address the class imbalance problem. SMOTE not only under samples the majority class, but oversamples the minority class by introducing ‘synthetic’ examples [14]. These examples are generated by adding a random number to the

difference between two neighbouring feature vectors of minority class examples. The amount of oversampling is a tuneable parameter.

3 Random Class Balancing

The class balancing ensemble classifier framework is presented in Figure 1. Given the training data a number of M subsets are produced using random under sampling. This is done to address the class imbalance problem. The under sampling process creates a training set with equal number of instances from every class. When the classifiers are trained on the balanced data set the bias mentioned above is eliminated. We term this method Random Class Balancing (RCB). The random selection process, however, may ignore significant clusters of the majority class. To address this issue we produce M different subsets from the training data using random under sampling. Classifiers are then trained on the M different training subsets.

During testing a sample is classified by all the M classifiers. The decisions produced by the M different base classifiers are merged into a single verdict using majority voting fusion. In this process the class that receives the maximum vote is considered to be the final classification verdict.

The class balancing ensemble classifier incorporates balancing to address the unfairness to the minority classes. Use of multiple subsets of the training data improves the overall sampling efficiency. Combined the accuracy of the minority class improves with minimal reduction of accuracy of the majority class.

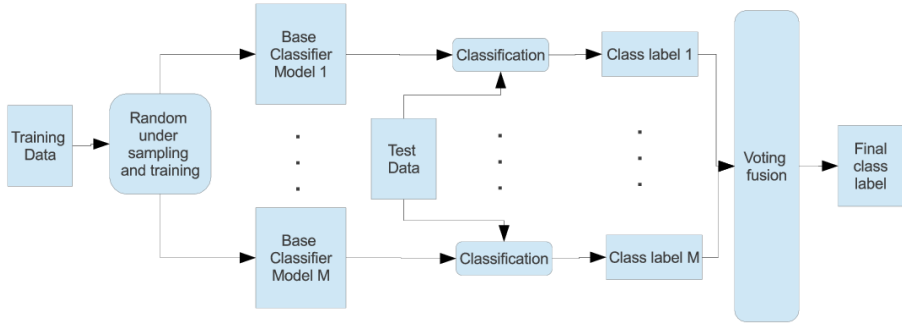


Fig. 1. Class Balancing Ensemble classifier framework

4 Dataset

The dataset is derived from 4543 manual water samples taken by TSQAP between 1988 and 2010. It includes four features:

- **Level of Thermotolerant Coliforms** - Level of faecal bacteria present per 100 millilitres.

- **Location** - 45 shellfish locations all over Tasmania.
- **Rain** - Rainfall in millimetres recorded for last 7 days from the closest weather station.
- **Salinity** - salt level in the water sample in Practical Salinity Units (PSU).

The sample dates have been merged with farm closure dates taken from TSQAP annual reports to create an output class, which is the close/open decision made by the TSQAP manager. We are particularly interested in the examples where the output class is *Closed* as this is when the farmers or the TSQAP manager will want to take action. Unfortunately, over 88% of the samples have the output class *Open*. However, we want to keep as many of the individual samples as we can, even if the farm was kept open, because as much information as possible is necessary to create an accurate model of these kinds of complex environmental processes.

5 Results

5.1 Feature Ranking

Determining which features are most relevant for predicting farm closures will inform our recommendations to farmers regarding what sensors to add. We also wanted to evaluate if learning could be simplified by removing any features that were irrelevant. We have used attribute/feature ranking as the search method to get an ordering of the features. An attribute evaluator method is required to evaluate the strength of an attribute. We have obtained evaluation scores of the attributes from eight different evaluators available in the Weka toolbox [15]. The scores and ranks are presented in Table 1 and Table 2. The average ranks of the attributes are obtained at the end to get an ensemble decision from all the evaluators.

From the average over the eight evaluators the following ordering of the attributes is observed, from most to least significant:

1. Location
2. Salinity
3. Rain
4. Coliforms

5.2 Class Balancing

The shellfish closure dataset was applied to four different classifier types from the Weka Toolbox [15]; Decision Trees (Table 3) with the J48 classifier, Random Forest (4), Multilayer Perceptrons (Table 5), Support Vector Machines (Table 6) using the SMO classifier, and Bayesian Networks (Table 7). The default parameters were used for all of the classifiers and 5-fold cross validation was performed on each. For the ensemble classifiers there were ten classifiers in each ensemble.

Table 1. Attribute score on the Shellfish Farm Closure dataset

Evaluator	Rain	Coliforms	Location	Salinity
Chi-squared Ranking Filter	316.3	307.4	901.3	429.2
Filtered Attribute Evaluator	0.045	0.042	0.1	0.06
Gain Ratio Feature Evaluator	0.02	0.34	0.037	0.04
Information Gain Ranking Filter	0.5	0.04	0.1	0.06
OneR Feature Evaluator	89.1	88.4	88.1	87.7
ReliefF Ranking Filter	0.03	0.001	0.03	0.01
SVM Feature Evaluator	4	3	1	2
Symmetrical Uncertainty Ranking Filter	0.03	0.05	0.07	0.06

Table 2. Attribute/Feature ranking on the Shellfish Farm Closure dataset

Evaluator	Rain	Coliforms	Location	Salinity
Chi-squared Ranking Filter	3	4	1	2
Filtered Attribute Evaluator	3	4	1	2
Gain Ratio Feature Evaluator	4	3	2	1
Information Gain Ranking Filter	3	4	1	2
OneR Feature Evaluator	1	2	3	4
ReliefF Ranking Filter	1	4	2	3
SVM Feature Evaluator	1	2	4	3
Symmetrical Uncertainty Ranking Filter	4	3	1	2
Average	2.5	3.2	1.9	2.8
Ranking	3	4	1	2

In Table 3 to 7 below, Random Class Balancing (RCB) refers to the average of the individual classifiers before they were added to the ensemble. For the Decision Tree tests we implemented three different voting methods. *Average* refers to the average vote over the ensemble, *Majority* is where a majority vote is taken with a random selection taking place on equal numbers of votes, and *Maximum* takes the vote of the classifier whose classification has the highest probability. The

Table 3. Accuracy of decision trees with ensemble and class balancing methods on the Shellfish Farm Closure dataset

Method	Overall %	Closed %	Open %
Default	91.9 ($\pm$ 1.8)	43.4 ($\pm$ 4.8)	98.0 ($\pm$ 0.9)
SMOTE	92.0 ($\pm$ 2.1)	73.5 ($\pm$ 3.6)	96.5 ($\pm$ 2.3)
Ensemble Avg	88.1 ($\pm$ 0.01)	0.0 ($\pm$ 0.0)	100.0 ($\pm$ 0.0)
Ensemble Maj	88.2 ($\pm$ 0.2)	35.0 ($\pm$ 48.7)	88.2 ($\pm$ 0.2)
RCB	78.8 ($\pm$ 2.5)	90.4 ($\pm$ 3.3)	77.2 ($\pm$ 5.1)
RCB Ensemble Average	80.1 ($\pm$ 1.6)	93.7 ($\pm$ 2.4)	78.3 ($\pm$ 1.7)
RCB Ensemble Majority	80.1 ($\pm$ 1.5)	98.9 ($\pm$ 2.5)	76.9 ($\pm$ 4.3)
RCB Ensemble Maximum	75.7 ($\pm$ 1.5)	93.8 ($\pm$ 2.5)	73.3 ($\pm$ 4.3)
AdaBoost	96.6 ($\pm$ 0.6)	82.0 ($\pm$ 4.0)	98.5 ($\pm$ 0.3)
Bagging	95.4 ($\pm$ 0.8)	72.4 ($\pm$ 4.8)	98.5 ($\pm$ 0.4)

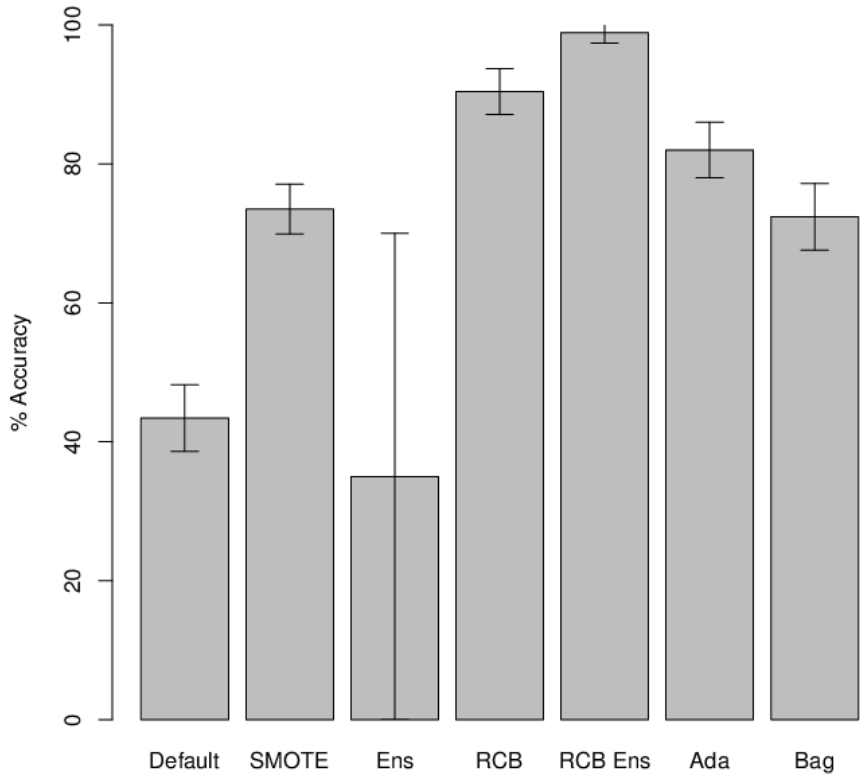


Fig. 2. Accuracy of the minority class, Closed, on the Shellfish Farm Closure dataset using Decision Tree classifiers

Table 4. Accuracy of Random Forest ensemble and class balancing methods on the Shellfish Farm Closure dataset

Method	Overall %	Closed %	Open %
Default	96.1 (± 0.4)	83.0(± 2.5)	97.8 (± 0.3)
SMOTE	95.0 (± 0.1)	83.5 (± 3.6)	96.5 (± 0.6)
RCB	87.6 (± 0.5)	91.6 (± 1.7)	87.0 (± 0.6)
RCB Ensemble	89.7 (± 0.5)	93.3 (± 1.7)	89.3 (± 0.6)

Table 5. Accuracy of Multi-Layer Perceptrons with ensemble and class balancing methods on the Shellfish Farm Closure dataset

Method	Overall %	Closed %	Open %
Default	87.8 (± 0.3)	46.8 (± 10.2)	98.0 (± 0.9)
SMOTE	85.3 (± 1.9)	28.1 (± 10.9)	93.0 (± 3.7)
RCB	67.2 (± 4.1)	72.1 (± 13.1)	66.5 (± 10.0)
RCB Ensemble	75.3 (± 4.1)	60.5 (± 3.8)	71.4 (± 10.3)

Table 6. Accuracy of SVM ensemble and class balancing methods on the Shellfish Farm Closure dataset

Method	Overall %	Closed %	Open %
Default	88.1 (± 0.01)	0.0 (± 0.0)	100.0 (± 0.0)
SMOTE	88.1 (± 0.01)	0.0 (± 0.0)	100.0 (± 0.0)
RCB	81.3 (± 2.8)	42.7 (± 5.2)	86.5 (± 3.6)
RCB Ensemble	81.6 (± 2.1)	45.2 (± 4.0)	86.6 (± 2.8)

Table 7. Accuracy of Bayes Net ensemble and class balancing methods on the Shellfish Farm Closure dataset

Method	Overall %	Closed %	Open %
Default	84.9 (± 0.6)	50.2 (± 3.3)	89.6 (± 0.7)
SMOTE	83.8 (± 0.6)	60.2 (± 3.6)	87.0 (± 0.7)
RCB	75.8 (± 1.4)	69.0 (± 3.3)	76.7 (± 1.7)
RCB Ensemble	75.6 (± 0.6)	70.2 (± 3.4)	76.3 (± 0.4)
AdaBoost	87.6 (± 1.6)	24.4 (± 13.7)	96.1 (± 3.2)
Bagging	85.4 (± 0.9)	46.1 (± 2.6)	90.7 (± 1.2)

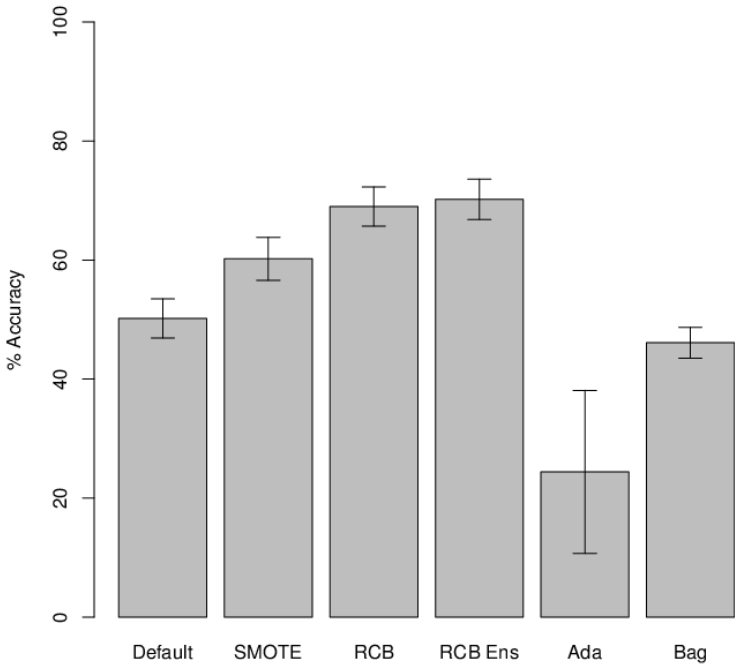


Fig. 3. Accuracy of the minority class, Closed, on the Shellfish Farm Closure dataset using Bayesian Network classifiers

remaining classifier methods use average voting as this is the default in Weka. The Bagging and AdaBoost classifiers were also taken from the Weka toolbox and used with the default parameters.

Figure 3 plots the accuracy on the *Closed* output class of interest for each of the decision tree approaches. Figure 7 plots the accuracy on the *Closed* output class of interest for each of the Bayesian Network approaches.

6 Discussion

The results of the feature ranking emphasise the importance of clustering the farms based on the location. It will not be possible to create a general classifier for every location, but we may be able to cluster the locations to avoid having a separate model for each location. Salinity is the most important environmental variable, followed by rain. It is interesting to note that the direct measures of thermotolerant coliforms are the least relevant when these are the levels we are directly trying to measure. It suggests that closures are carried out in anticipation of drops in thermotolerant coliforms most often, rather than due to high levels of the coliforms themselves. This will be important for our real-time system as the health department and the farmers will want to be forewarned of potential closures so, for example, they can harvest early, instead of closing when it is essentially too late to take appropriate action.

The results of the learning methods show that the classification accuracy can be significantly improved by class balancing techniques. SMOTE shows an improvement over the default method for Decision Trees, Random Forest and Bayesian Networks on the minority class; with the largest increase in accuracy apparent for Decision Trees. The Random Forest proving the best performer using the default method.

The Random Class Balancing Ensembles consistently had the highest accuracy on the minority *Closed* output class. Although it did negatively affect the accuracy on the entire data set and the *Open* output class. We will need to investigate if there is a method that does not decrease the accuracy of the majority class when class balancing. The results suggest that SMOTE also loses accuracy in the majority class. The Bagging and Boosting were able to greater overall accuracy, but significantly lower accuracy on the minority class than the RCB Ensemble method. The Bagging and Boosting methods also had higher variance in accuracy.

There was a small increase in accuracy using majority voting, but similar results were gained from the average and maximum voting methods. For the remaining experiments we left the voting as the average voting default as the accuracy increase was not deemed significant enough.

The accuracy of MLPs and SVMs likely suffer from the format of the location, which has been determined as the most important feature. The location information would not translate well to a single normalised input. It would also not be feasible to have an input node for each of the 45 locations. We believe the neural network learners will respond well to reducing the locations to a much smaller

number of clusters before learning. However, these learners still experience a significant increase in accuracy after class balancing.

7 Conclusion

We were successfully able to train classifiers on the Shellfish Farm Closure dataset and produce accurate classifications of farm closures. The application of machine learning techniques in this domain is novel in itself. The classifiers were consistently improved by using class balancing techniques; with 98.9% accuracy on the minority class possible with the RCB Ensemble method described. We also confirmed the importance of salinity as a indictator for farm closure, as well as the importance of the farm location.

We intend to make these classifications with input from real-time data streams to provide an intuitive visualisation tool for the health authority. The probability of closure will be represented by a traffic light colour system on a geographic map. We believe this greatly reduce the workload of the manager as well as the current dependence on her skills.

Acknowledgements. This work is supported by the Food Futures Flagship, CSIRO and the Tasmanian State Government Department of Economic Development, Tourism and the Arts.

References

1. Muttill, N., Chau, K.: Machine-learning paradigms for selecting ecologically significant input variables. *Journal Engineering Applications of Artificial Intelligence* 20(6), 735–744 (2007)
2. Rahman, A., Verma, B.: Novel layered clustering-based approach for generating ensemble of classifiers. *IEEE Transactions on Neural Networks* 22(5), 781–792 (2011)
3. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
4. Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
5. Breiman, L.: Pasting small votes for classification in large databases and on-line. *Machine Learning* 36, 85–103 (1999)
6. Martinez-Munoz, G., Hernandez-Lobato, D., Suarez, A.: An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31(2), 245–259 (2009)
7. Chen, L., Kamel, M.: A generalized adaptive ensemble generation and aggregation approach for multiple classifiers systems. *Pattern Recognition* 42, 629–644 (2009)
8. Nanni, L., Lumini, A.: Fuzzy bagging: a novel ensemble of classifiers. *Pattern Recognition* 39, 488–490 (2006)
9. Eschrich, S., Hall, L.O.: Soft partitions lead to better learned ensembles, pp. 406–411 (2002)
10. Schapire, R.: The strength of weak learnability. *Machine Learning* 5(2), 197–227 (1990)
11. Freund, Y., Schapire, R.: Decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1), 119–139 (1997)

12. Drucker, H., Cortes, C., Jackel, L., LeCun, Y., Vapnik, V.: Boosting and other ensemble methods. *Neural Computation* 6(6), 1289–1301 (1994)
13. Garcia-Pedrajas, N.: Constructing ensembles of classifiers by means of weighted instance selection. *IEEE Trans. on Neural Networks* 20(2), 258–277 (2009)
14. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 341–378 (2002)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. *SIGKDD Explorations* 11(1) (2009)