Computers and Electronics in Agriculture 110 (2015) 241–248

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Shellfish farm closure prediction and cause identification using machine learning methods

Ashfaqur Rahman *, Claire D'Este

Autonomous Systems, Digital Productivity & Services Flagship, CSIRO, College Road, Sandy Bay 7005, Australia

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 17 November 2014
Accepted 22 November 2014

Keywords:
Ensemble feature ranking
Prediction
Shellfish farm closure prediction

ABSTRACT

Shellfish farms are needed to be closed if they are contaminated during their production as otherwise it may lead to serious health hazard. The authorities monitor a number of water quality variables to check the health of shellfish farms and to decide on the closure of the farms. The research presented in this paper aims to automate this process by developing novel algorithms to identify the cause of closure and also predicting the closure. As the frequency of closure is relatively very small, the labelled data sets are imbalanced in nature. We have developed a novel ensemble feature ranking algorithm that explicitly deals with class imbalance problem and identifies the cause of closure. We have also presented a class balancing ensemble classifier to predict shellfish farm closure. The class balancing ensemble classifier predicts closure/opening with as high as 71.69% accuracy and achieves best balancing act with decision tree base classifier in 75% locations. Rain and salinity are found to be the key causes of closure and the causality depends of the properties of the locations.

1. Introduction

Consumption of contaminated shellfish can cause severe illness and even death in humans. Authorities such as the Tasmanian Shellfish Quality Assurance Program (TSQAP) are responsible for monitoring the microbial contaminant levels of commercially grown shellfish. These organisations have the authority to close farms if they are concerned by the water quality at the site. 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 local industry.

Decision support systems (DSS) aim to assist in managerial decision-making, but not to replace the manager. However, capturing the full operational knowledge of the expert is notoriously difficult even for the expert himself or herself. In this research we deal with a data set of environmental samples taken from manual water samples and weather stations. We also have access to the dates of closures for each of the growing zones. These two data sets have been combined to provide the environmental context and the actual decision made. This provides a labelled data set from which to learn the conditions that suggest a closure should be made. Individual thresholds are currently created for each location from collecting manual samples and performing linear regression on salinity and rainfall with respect to thermotolerant coliforms. Thermotolerant coliforms represent the microbial contaminants that pose a risk to public health. Through data sets, such as the one investigated in this paper, we aim to capture the decision making process of the TSQAP manager using data-driven approaches, which may include unconscious processes, and potentially more complex processes than simple thresholds. There is also the potential to create generalized models that can be applied in new growing locations, or where manual sampling has not provided sufficient ranges of coliform levels with which to build an accurate model.

We have developed novel machine learning algorithms to identify the cause of shellfish farm closure and predict opening/closure. The data set (to be used by the machine learning framework) is composed of environmental and water quality variables as features and the farm opening status (open/close) as the class feature. The farms are generally closed for a shorter period compared to the period it is open. This results in an imbalanced data set with 'Close' being the minority class. We have developed a novel ensemble feature ranking algorithm to identify the cause of farm closure. The ranking algorithm features the capability to handle imbalanced data sets. We have also developed a mechanism to associate the characteristics of a location to the cause of closure using joint

* Corresponding author.
E-mail addresses: ashfaqur.rahman@csiro.au (A. Rahman), claire.d'este@csiro.au (C. D'Este).

http://dx.doi.org/10.1016/j.compag.2014.11.023
0168-1699/

probability distribution table. We have also developed a novel class balancing ensemble classifier to predict shellfish farm closure. We have utilized the above frameworks to answer the following research questions: (i) what is the prime cause of shellfish farm closure, (ii) how does location characteristics associate with the cause of closure, and (iii) how well can we predict the shellfish farm closure.

The paper is organized as follows: Section 2 presents some related works on aquaculture decision support systems, feature ranking and ensemble classifier algorithms. Section 3 presents the proposed ensemble feature ranking and classification framework. The underlying framework used in the experiments is presented in Section 4. Section 5 presents the findings and the analysis on the results. Finally, Section 6 concludes the paper.

2. Related works

Generic decision support systems have been developed for aquaculture farm operations. Bourke et al. (1993) facilitate aquaculture research by the display of real-time water quality indicators, as well as operational information, such as stock density and initial size, to evaluate their impact on survival rate, biomass and production failure using an eigenvalue method. Wang et al. (2006) encode heuristics to form an early warning system for dangerous growing conditions. Similarly Padala and Zilber (1991) use rules derived directly from an expert to reduce stock loss and increase size and quality of yields. Ernst et al. (2000) focus on managing hatchery production; using quantitative methods and models gained from an aquatic chemist, aquacultural engineer, aquatic biologist and fish biologist. The methods include rules and calculations of physical, chemical and biological processes.

Decision support systems are also developed for aquaculture to inform farm site selection. Silvert (1994) developed a DSS for evaluating the environmental impact of potential farms using scientific models. Halide et al. (2009) approach site selection from the perspective of its economic performance; where again the rules are hand-crafted from domain experts.

None of the decision support systems described above use a data-driven approach for developing the models. Data mining/machine learning techniques are rarely applied to aquaculture problems; with one exception being the prediction of harmful algal blooms (Muttil and Chau, 2007). Shellfish farm closure prediction remains to date a novel application of data-driven techniques.

We have presented a new feature ranking algorithm and an ensemble classifier in this paper. For the sake of completeness we present here some commonly used feature ranking algorithms and ensemble classifiers. Feature selection algorithms (Tsang et al., 2001; Yeung and Wang, 2002; Jarmulak and Craw, 1999; Liu and Kender, 2003; Wettschereck et al., 1997) that can be classified into wrapper and filter methods. Wrapper methods evaluate features by seeking feedback from subsequent classifier whereas filter methods assess the relevance of features (Sayes et al., 2007; Rahman and Murshed, 2004) on the basis of the intrinsic properties of the data. The filter methods provide a ranking of the features in terms of their relevance to the classes and in the research presented in this paper we focus on filter methods. Feature ranking algorithms evaluate attributes using Information gain, OneR classification, SVM classification, Gain ratio, Chi squared test, and Symmetrical uncertainty (Hall et al., 2009). The proposed ensemble feature ranking algorithm obtains ranking of features using these existing methods and integrates these results into a cumulative ranking. The novelty in the proposed ensemble ranking algorithm lies with the aggregation of the base feature ranking algorithms.

An ensemble classifier refers to a group of classifiers trained simultaneously to obtain better performance than their base counterparts. An ensemble classifier performs better than its base counterparts if the base classifiers are accurate and diverse. Diversity (Tsang et al., 2001) refers to the complementary nature of learning of the base classifiers. Training data is manipulated (Rokach et al., 2003; Breiman, 1996; Rahman and Verma, 2011; Rahman et al., 2010; Schapire, 1990; Freund and Schapire, 1997) to obtain diversity among the base classifiers. Bagging (Breiman, 1996) is a commonly used ensemble classifier where multiple training subsets are generated randomly and identical base classifiers are trained on the subsets. The class chosen by the majority base classifiers is the verdict of the ensemble classifier. In (Rahman and Verma, 2011; Rahman et al., 2010) data is clustered into multiple layers and classifiers are trained on clusters at each layer. A pattern is classified at each layer by the classifier trained on the nearest clusters and the decisions from all the layers are fused into a single verdict using majority voting. In Boosting (Schapire, 1990) each training example is assigned a weight that determines how well the instance was classified in the current iteration. The training data that are misclassified in the current iteration are assigned higher weight for inclusion in the training subset for the next iteration. AdaBoost (Freund and Schapire, 1997) is a more generalized version of boosting. None of the above ensemble classifiers consider the class imbalance problem while constructing the base classifiers. We present a class balancing ensemble classifier to deal with this problem. A preliminary version of this work can be found at (D'Este et al., 2012).

3. Proposed methods

We have expressed shellfish farm closure prediction as a supervised learning problem. We have identified variables used by farmers to decide on farm closure and obtained their historical readings. We have also obtained the opening status (Open/Close) of the shellfish farms at corresponding times. We constitute a data set combining the readings of environmental variables as input and farm closure status as class/output. Cause of a shellfish farm closure is identified by obtaining the ranking of the features in terms of their relevance to closure/opening. Prediction of closures is obtained by training classifiers. The data set however is imbalanced by nature. It is thus necessary to apply balancing actions to refrain from obtaining a misleading ranking and also to unfairly treat the minority class. The following sections present a novel ensemble approach to obtain feature ranking and perform classification that has an inherent mechanism to deal to class imbalance.

3.1. Class balancing through random under sampling

Let the training data be denoted by Y such that

$$Y = [X \quad Q] \tag{1}$$

where X is the data matrix and Q is the class vector. $[X_i\ Q_i]$ represents the training subset containing the instances corresponding to class i where $i \in \{1, \ldots, N_q\}$ and N_q is the total number of classes. Let w_i represent the number of rows (i.e. instances) in matrix $[X_i\ Q_i]$ where $i \in \{1, \ldots, N_q\}$. The number of samples corresponding to minority class is identified as

$$\eta = \text{argmin}_{i \in \{1,\ldots,N_q\}} \omega_i \tag{2}$$

A random under sampling function $\Re$ samples η instances from all the training subsets $D_i = [X_i\ Q_i]$ randomly where $i \in \{1, \ldots, N_q\}$. The random under sampling process forms the basis for the ensemble feature ranking and ensemble classification process as detailed next.