Engineering Applications of Artificial Intelligence 34 (2014) 58–65

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Ensemble aggregation methods for relocating models of rare events

Claire D'Este[a,*], Greg Timms[a], Alison Turnbull[b], Ashfaqur Rahman[a]

[a] *CSIRO Computational Informatics, Castray Esplanade, Hobart 7000, Australia*
[b] *Department of Health and Human Services, GPO Box 125, Hobart 7001, Australia*

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form
2 April 2014
Accepted 8 May 2014
Available online 4 June 2014

Keywords:
Ensemble classifiers
Rare event detection
Aquaculture

ABSTRACT

Spatially distributed regions may have different influences that affect the underlying physical processes and make it inappropriate to directly relocate learned models. We may also be aiming to detect rare events for which we have examples in some regions, but not others. Three novel voting methods are presented for combining classifiers trained on regions with available examples for predicting rare events in new regions; specifically the closure of shellfish farms. The ensemble methods introduced are consistently more accurate at predicting closures. Approximately 63% of locations were successfully learned with Class Balance aggregation compared with 37% for the Expert guidelines, and 0% for One Class Classification.

1. Introduction

Creating a real-time decision support system to detect rare events from sensor data streams using expert knowledge suffers from the well-known problem that experts find explaining their decision process extremely difficult. When monitoring unreliable, noisy sensor data streams experts are often including additional features, extending temporal and spatial windows, filling in missing values, and mentally adjusting absolute values based on other context. This kind of workflow is difficult to capture in simple heuristics.

The Aquaculture Decision Support (AquaDS) project has been focusing on developing a real-time decision support system for the Tasmanian Shellfish Quality Assurance Program (TSQAP) to reduce the dependence on a single expert. The systems produce nowcast and forecasts of shellfish farm closures over the web to allow farmers at 38 growing zones to manage proactively. Currently closure guidelines are applied on a zone-by-zone basis because of the effect of geography on the physical processes as well as differences in distance and availability of sensors.

The AquaDS project is developing data-driven models to predict closures from a database of manual samples linked to closure dates. The closures, however, are relatively rare with most growing zones having less than 10% manual samples taken when they were closed, Table 1. Although we have demonstrated that this problem can be learned with standard machine learning techniques using class balancing techniques, such as undersampling (D'Este et al., 2012) there are still many locations where we have insufficient, or no, examples of closures. Delivering a decision support system that predicts for only a small number of growing zones will not be satisfactory. There is also a high probability in the future that new growing zones will open and it would be beneficial to begin providing closure predictions as soon as possible.

This paper focuses on methods for predicting rare events for multiple regions that are affected by their geographical location. The approach uses a combination of classifiers from the other zones whose classifications on the *Open*/*Closed* state are weighted by three novel methods using Matthews Correlation Coefficient accuracy, the Similarity of the relationship between water quality and coliform levels, and the Class Balance of the training set. These are compared with Expert guidelines, One Class Classification and ensembles with traditional average, maximum probability and classification accuracy aggregation methods.

2. The shellfish farm closure problem

Farmed shellfish contaminated with harmful microbes can cause serious risk to public health. Areas where shellfish are grown in Australia, and many other parts of the world, are monitored to ensure product does not make it to market that might cause serious illness or death when consumed. The TSQAP manager must monitor farms spread out over a 68,401 square kilometre area and relies on a variety of tedious manual processes to gain water quality information in each growing area. The water quality information is extracted from a range of organisations via a range of means, including dial-up ftp servers, web pages and phone calls to farmers. The farmers themselves desire the shortest

[*] Corresponding author.
E-mail address: claire.deste@csiro.au (C. D'Este).

http://dx.doi.org/10.1016/j.engappai.2014.05.007
0952-1976/Crown Copyright © 2014 Published by Elsevier Ltd. All rights reserved.

possible closure times, which costs them around US $5000 for each day they cannot harvest and may last for months. Long closure times can also result in significant stock loss and loss of market share. As well as aggregating real-time information from sensors from all the necessary organisations, the AquaDS project will also provide predictions of the current probability of closure. The TSQAP program currently relies heavily on the expertise of a single manager, which can result in missing risky conditions when she is unavailable. Therefore, she has developed a list of closure guidelines that can potentially be used by other members of the health department in her absence. The guidelines are hard thresholds designed to fit on a small spreadsheet that can be referred to by someone who is untrained. We hypothesise that, although the manager is highly skilled at closing farms appropriately, these rules do not adequately represent her decision making process.

The closure guidelines have been developed based on the response of thermotolerant coliforms (faecal bacteria) to changes in salinity, rainfall and river flow. The salinity, rainfall and river flow are then monitored remotely to ensure that they have not exceeded thresholds.

The thresholds themselves are developed by plotting the environmental phenomenon (salinity, rainfall or river flow) against the thermotolerant coliform levels. The expert then determines the point at which the coliform level rises to unsafe level. The authorities use some base statistics, the median must be below 14 coliforms per 100 ml and 90% of samples must be under 21.

Table 1
Number of examples and percentages of each open/closed class for the growing zones studied, and marker number in Fig. 2.

Zone name	Marker	Closed %	Open %	Instances
Big Bay B	1	32.5	67.5	209
Big Bay C	1	22.2	77.8	266
Big Bay E	1	21.1	78.9	665
Blackman	2	11.7	88.3	729
Blackman Bay East	2	6.7	93.3	419
Blackman Bay LBB	2	0	100	93
Cloudy Bay	3	4.0	96.0	199
Deep Bay	4	15.4	84.6	415
Duck Bay	5	35.5	64.5	900
Dunalley A	6	8.8	91.2	495
Dunalley B	6	0	100	39
Eaglehawk Bay	7	0.8	99.2	494
Fleurty's Point	8	9.1	90.1	484
Gardner's Bay	9	0	100	417
Garfish Bay	10	0	100	412
Great Bay	11	2.0	98.0	809
Great Oyster Bay	12	3.2	96.8	377
Great Swanport West	13	3.1	96.9	819
Great Swanport East	13	0	100	103
Hastings	14	15.8	84.2	941
Island Inlet 4	15	0	100	586
Island Inlet 5	15	0	100	116
Little Swanport	16	3.1	96.9	879
Little Taylors Bay	17	98.1	1.9	318
Montagu	18	23.5	76.5	337
Moulting Bay 1	19	11.7	88.3	367
Moulting Bay 2	19	0	100	658
Moulting Bay 4	19	0	100	555
Moulting Bay 5	19	0	100	565
Moulting Bay 6A	19	0	100	466
Norfolk Bay	20	0	100	276
Pipe Clay 1	21	6.5	93.5	897
Pipe Clay 3	21	7.1	82.9	546
Pitt Water 1	22	2.3	97.7	1145
Pitt Water 2	22	0	100	18
Pitt Water 3	22	0	100	937
Port Arthur	23	5.1	94.9	217
Port Esperance 1A	24	0	100	93

Fig. 1(a) demonstrates this method for salinity at Fleurty's Point. We can visualise that the percentage of coliforms over 21 per 100 ml start rising when salinity drops below 31 PSU. The actual salinity threshold is <30 PSU for this zone. Great Swanport West however Fig. 1(b) shows risky coliform levels rising when salinity drops under 25 PSU, which is consistent with the actual salinity trigger of <26.

The manager performs outlier detection and adjustments based on her expertise. For example, accounting for sensor drift on various platforms.

3. Related work

A broad range of decision support systems for a variety of aquaculture purposes are currently available. Their uses range from selecting and licensing aquaculture sites (Silvert, 1994a, 1994b; Stagnitti, 1997; Stagnitti and Austi, 1998), facility design and production planning (Ernst et al., 2000), planning nutrient removal facilities (Vezzulli et al., 2006), managing hatchery production (Schulstad, 1997), forecasting aquaculture production (Zhang et al., 2005; Menicou et al., 2010), predicting environmental impact under various management scenarios (Casini et al., 2005; Conte and Ahmadi, 2010), facilitating aquaculture research and management (Bourke et al., 1993) and performing economic impact evaluation (Bolte et al., 2000).

This work is generally planning or scenario-based; it does not seek to supply nowcasts or forecasts and/or recommend operational decisions based on real-time data or short-term predictions. In general, data mining/machine learning techniques are rarely applied to aquaculture problems, with the exception of prediction of harmful algal blooms (Muttil and Chau, 2007).

Work related to the AquaDS project was undertaken by Chigbu et al. (2006). They developed a system for regulators that integrated rainfall and streamflow data from the National Oceanic and Atmospheric Administration (NOAA) and the National Weather Service (NWS), and compared with regulated thresholds to recommend a close/open decision. Their work did not seek to ingest feedback from decisions, make predictions and/or address the needs of the farmers themselves.

Kelsey et al. (2010) investigated the assumed relationship between rainfall/streamflow and faecal coliforms. They performed a multiple-parameter regression analysis over a much wider range of historical environmental and water sampling data in four US estuaries to identify whether closure decisions would be better based on other proxies. They point out that “These … could be used to develop real-time predictions of bacteria concentration for use in a closure decision system. Rainfall measurements at gauges or from NEXRAD are available in near-real time through the NOAA NWS, and it may be relatively straightforward to develop real-time data sources for salinity and water temperature as well.”

Wang et al. (2006) outlined a system for forecasting water quality in an aquaculture pond using an expert system's approach, however do not appear to have advanced this work and made their system operational.

Rare event detection is generally performed using class imbalance techniques that alter the training set using oversampling of positive examples and/or undersampling of negative examples (Chawla et al., 2002). However, these still require positive examples to be available.

One Class Classification (Tax, 2001) provides the ability to learn in the absence of examples of the target class. A model is created on the negative examples and outliers are presumed to represent the target class.

Minku and Yao (2012) use the output of different software development companies to predict the productivity of another