ИСПОЛЬЗОВАНИЕ МНОГОУРОВНЕВОЙ МОДЕЛИ ДЛЯ ЭФФЕКТИВНОГО УПРАВЛЕНИЯ ДАМБОЙ И ПРЕДСКАЗАНИЯ НАВОДНЕНИЙ

Шлеменков А. А., Гусак Я. О.
Факультет компьютерных систем и сетей, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {alex.shlemenkov, yaninagusak}@gmail.com

В данной работе будет предложена многоуровневая модель, которая значительно улучшает качество предсказаний потока в дамбе по сравнению со стандартными подходами.

Введение

Наводнение – одно из наиболее разрушительных стихийных бедствий. Для минимизации вероятности необратимых последствий необходимо не только как можно более точно предсказывать поток в дамбе, но и делать это заранее. Такие предсказания помогут скорректировать управление дамбой и избежать катастрофы. Другой целью моделирования служит оптимизация производства электроэнергии дамбой.

I. Описание данных

Данные задачи представляют собой временные ряды со значениями потока через дамбу в текущий момент, количеством осадков и уровнем воды в метрах на нескольких мостах вверх по течению реки. Также для анализа доступны уровни осадков в мм на площади в 900 км^2. Эта область почти полностью покрывает бассейн реки, на которой располагается дамба. Форма региона визуализирована на рисунке ниже.

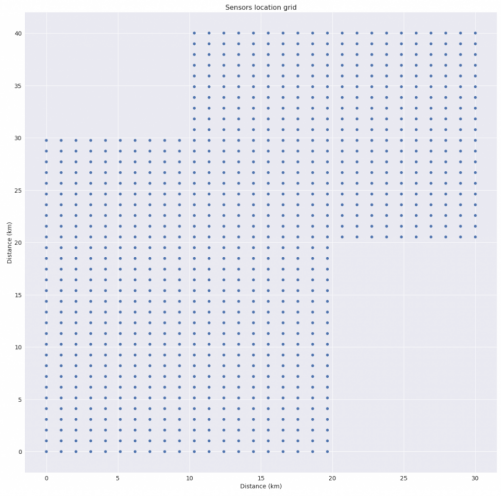

Рис. 1 – Форма области с датчиками

Некоторые виды данных из упомянутых выше не были полными: что-то отсутствовало случайно, у некоторых показателей пропуски были периодическими, поток в дамбе был отрицательным для некоторых промежутков времени. Убирая из рассмотрения слишком неполные или шумные события, нам удалось получить около 20000 точек и 50 событий. Пример одного из таких событий указан на рисунке.

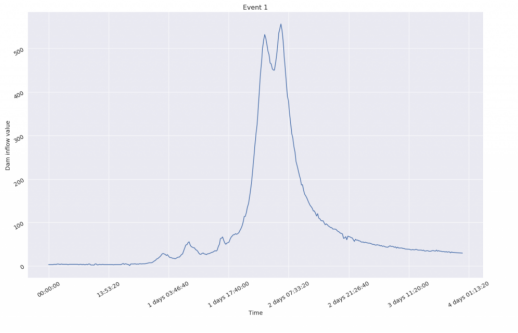

Рис. 2 – Пример экстремального события

II. Постановка задачи

Формальная постановка задачи звучит следующим образом: предсказать поток в дамбе через 3 часа. Трехчасовой период времени был выбран из-за того, что этого достаточно для оперативного вмешательства в управления дамбой для предотвращения наводнения. В качестве метрики оценки качества модели предлагается RMSE. В качестве метрики для оптимизации использоовалась функция RMSLE [1], которая является хорошим приближением RMSPE, но, в дополнение, еще и хорошо дифференцируется. RMSPE позволяет модели оптимизировать относительную ошибку, а не абсолютную. Т.е, например, для ошибки в предсказании величиной 10 пунктов при реальном значении в 200 и 20 путнков соответственно, штраф модели не будет одинаковым. Для линейных моделей выбор функционала ошибки является очень важным шагом и сильно влияет на качество предсказаний.

III. Техника валидации

В качестве техники валидации моделей была выбрана leave-one-out [2] схема, в которой обучение производится на всех кроме одного события, а после оценивается на этом отложенном событии. Дополнительной мотивацией для выбора

такой техники была высокая степень подобия с использованием модели в реальном мире.

IV. Моделирование

1. Для получения базового решения можно использовать линейные модели. В качестве признаков использовались значения всех доступных данных за последние 6 часов. Этого шестичасового окна было лостаточно для понимания динамики. Полученные результаты дали среднее значение по схеме leave-one-out в размере 21.12 по метрике RMSE.
2. Другой идеей было использование сверточных сетей для анализа области в 900 км^2, но, к сожалению, нейросеть не смогла показать достаточную предиктивную силу. После визуального анализа осадков в данной области, не было заметно какой-либо корреляции между потоком и осадками. Низкое качество признаков–осадков подтверждает и тот факт, что коэффициент корреляции между суммой осадков в области и целевой переменной оказался значительно меньше, чем корреляция между текущим значением потока и целевой переменной: 0.6 и 0.825 соответственно.
3. Также была протестирована идея использования RNN, но основным, по нашему мнению, недостатком данных моделей была требовательность к объему данных. Несмотря на то, что GRU намного лучше показывает себя используя меньшее их количестно, качество модели оказалась хуже базового, полученного линейной моделью.
4. Использование ансамбля решающих деревьев [3] показало себя хорошо на нашем наборе данных. В терминах величины RMSE результаты были немного лучше базовой модели. Несмотря на это, у методов, которые основаны на решающих деревьях, есть большой недостаток – их невозможность экстраполяции, а в задаче это является очень критичным требованием.

V. Обучение многоуровневой модели

После большого количества экспериментов, было принято решение реализовать комбинированный подход: использовать градиентный бустинг над решающими деревьями на низком значении потока, а линейные модели на высоком значении потока для экстраполяции значения потока. На выбор многоуровневого подхода к финальной модели повлияло и количество данных: малый поток был представлен намного чаще, чем большой. Для того, чтобы использовать трюк с двумя моделями мы обучили SVM [4], который бы разделял точки на два класса: малый поток и большой поток. Точки из каждой группы подавались в свой регрессор. Конкретное разделяющее значение для малого и большого потока было найдено через множественную процедуру кросс-валидации многоуровневой модели. Для получения конечной модели мы проделали следующие шаги:

1. Разбили данные на N событий.
2. Обучили N SVM классификаторов на N-1 разбиении каждый и предсказали для всех точек на оставшемся событии уровень потока (низкий или выскокий).
3. Обучили градиентный бустинг на точках, которым был предсказан низкий уровень потока.
4. Обучили линейную регрессию на точках, которым был предсказан высокий уровень потока.
5. Обучили SVM класификатор на всех N событиях.

VI. Результат

Основая идея подхода – это создание многоуровней модели из двух моделей–специалистов, которые хорошо проявляют себя в своей области. Такая техника позволила снизить ошибку с 21.12 до 10.87 по метрике RMSE. Сравнение с базовой моделью показано на рисунке. На нем видно, что финальная модель реагирует быстрее на выбросы и предсказывает, в целом, точнее.

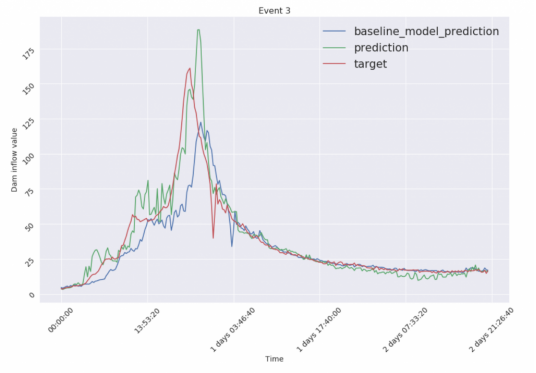

Рис. 3 – Сравнение базовой и многоуровневой моделей

1. RMSLE cost function [Electronic resource]. – https://www.slideshare.net/KhorSoonHin/rmsle-cost-function Date of access: 20.09.2018.
2. Evaluation: Leave One Out Cross Validation [Electronic resource]. – Mode of access: https://www.coursera.org/lecture/predictive-analytics/evaluation-leave-one-out-cross-validation-DfEWO – Date of access: 20.09.2018.
3. LightGBM [Electronic resource]. – Mode of access: https://github.com/Microsoft/LightGBM – Date of access: 20.09.2018.
4. Support Vector Machine – Theory [Electronic resource]. – Mode of access: https://goo.gl/mTBnV8 – Date of access: 20.09.2018.