В настоящее время технологии машинного обучения применяются для решения очень широкого спектра задач. Их использование для обработки, анализа информации и построения математических моделей на основе данных дистанционного зондирования Земли может поспособствовать разработке новых программных инструментов и улучшению автоматизации все большего спектра операций.

Использование технологий машинного обучения подразумевает, что проблема решается не традиционным написанием компьютерной программы для решения какой-либо задачи, а обучением уже готовых алгоритмов на частных примерах ее решения. Такой подход позволяет выполнить автоматизированное построение математической модели, подходящей для практических всех возможных вариантов исходных данных и вариантов решения рассматриваемой задачи. Авторами, по результатам применения технологий машинного обучения для конкурсных (GeoHack.112, Raiffeisen Data Cup, DengAI: Predicting Disease Spread, Statoil/C-CORE Iceberg Classifier Challenge) и научно-исследовательских работ (прогнозирование развития городских агломераций, автоматизированная векторизация гидрографических данных по спутниковым снимкам), ориентированных на обработку данных дистанционного зондирования Земли, представленных в исходном, либо обработанном виде, были сформулированы типовой порядок действий (приведен на рисунке) и рекомендации по применению алгоритмов в зависимости от типа решаемой задачи, приведенные далее в тексте статьи[1].

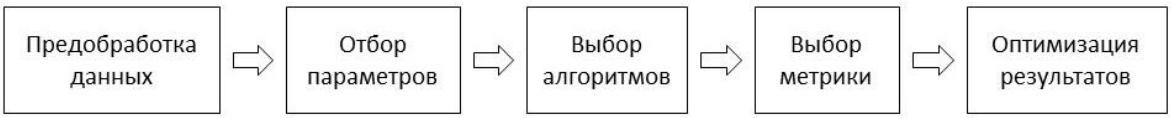

Рисунок – Типовая схема действий при создании математической модели с использованием технологий машинного обучения

Поскольку обучение программы происходит на основе набора данных (англ. dataset), то их предварительная обработка (предобработка) является очень важным этапом. Ошибки в типах данных, пропуски, шумовые значения и тому подобное могут очень сильно исказить итоговую математическую модель. На этом этапе проводится проверка и назначение (при необходимости) правильных типов параметров (целое или дробное число, текст, дата и время и т.п.), заполнение пропущенных значений (если это возможно и в этом есть необходимость, поскольку некоторые алгоритмы не могут обрабатывать данные с пропусками), удаление ошибочных данных, различные типы кодирования. Под кодированием здесь понимается замена исходных значений на те, которые больше подходят для выбранного алгоритма. Примерами этой операции могут быть конвертация даты и времени в набор целых чисел, отображающих количество недель, дней, часов от определенной даты или замена числового значения класса объекта на набор отдельных признаков все из которых равны нулю за исключением одного (англ. One-Hot Encoding).

После предобработки данных нужно выполнить операции по отбору тех параметров, по которым будет строиться итоговая модель. Данный этап необходим потому, что слишком большое число параметров может значительно замедлить выполнение алгоритмов и, что более важно, уменьшить точность предсказания и привести к переобучению (особенно при наличии параметров с низкой степенью корреляции с целевой переменной). Способами провести отбор значимых параметров могут быть:

1. визуализация с помощью диаграмм, графиков и т.п., в том числе на карте, посредством геоинформационных систем;

2. расчет значений корреляции (коэффициенты Пирсона, Спирмена и т.д.) [2-4];
3. пространственная автокорреляция, хаотичность размещения объектов в пространстве (индекс Морана и т.д.) [2-5];
4. расчет энтропии, в том числе для временных рядов (характерный показатель Ляпунова, коэффициент Хёрста, detrended fluctuation analysis и т.д.) [2-4].

Третий и четвертый пункты приведенного списка относятся к анализу пространственно-временных данных, в том числе данных ДЗЗ. Их расчет реализован как в традиционных геоинформационных системах (ArcGIS, GRASS GIS, Saga), так и в специализированных программных библиотеках (PySAL, nolds и других).

После того как был произведен отбор параметров нужно осуществить выбор наиболее подходящих алгоритмов. На данном этапе трудно однозначно сделать выбор в пользу того или иного алгоритма или технологии машинного обучения, поэтому обычно производят тестирование нескольких алгоритмов и уже потом производят оптимизацию для наиболее подходящего. Большое количество предложений и относительная дешевизна облачных сервисов для построения моделей машинного обучения позволяют запустить оценку точности одновременно для нескольких алгоритмов/технологий, по сути осуществить выбор оптимального и наиболее точного варианта методом подбора. Но такой вариант подходит не для всех задач, учитывая традиционно большой объем данных дистанционного зондирования. Далее приведен список наиболее часто используемых алгоритмов для конкретного круга задач:

- детекция и сегментация объектов на растровых данных [6-10];
  - деревья решений;
  - комбинация деревьев решений и градиентного бустинга (данный вариант является одним из наиболее универсальных решений, наряду с нейронными сетями);
  - традиционные и сверточные нейронные сети (являются наиболее универсальным решением, но требуется подбирать большое число параметров и подбирать архитектуру сети);
- классификация данных;
  - деревья решений и случайный лес;
  - нейронные сети;
  - метод опорных векторов;
  - логистическая классификация;
  - наивная байесовская классификация;
- расчет пропущенных. либо недостающих значений;
  - деревья решений и случайный лес;
  - комбинация деревьев решений и градиентного бустинга;
  - линейная регрессия;
  - логистическая регрессия;
  - нейронные сети;
  - ridge регрессия;
  - lasso регрессия;
  - метод опорных векторов;
  - кригинг и кокригинг (этот и следующие методы реализованы в большинстве геоинформационных систем);
  - метод обратных взвешенных расстояний [7];
  - методы локальных и глобальных полиномов;
- прогнозирование и анализ временных рядов (для этих задач характерно наличие пространственных и атрибутивных данных для одной и той же территории в нескольких временных интервалах) [8,10];

- комбинация деревьев решений и градиентного бустинга
- ARIMA/SARIMA;
- нейронные сети глубокого обучения;
- нейронные сети с долгой краткосрочной памятью (LSTM);
- пространственно-временные повторяющиеся сверточные сети (SRCN);
- пространственный анализ методом Монте-Карло (SMCA);

- поиск аномалий в данных;
  - метод опорных векторов;
  - метод главных компонент;
  - метод k-средних.

Для оценки результатов работы алгоритмов необходимо выбрать способ оценки качества и эффективности конкретного алгоритма (метрика) и их выбор также необходимо делать, основываясь на особенностях решаемой задач. При обработке результатов, имеющих среди исходной информации данные дистанционного зондирования можно использовать специфические метрики, учитывающие пространственное положение объектов - коэффициент Джини, коэффициент Кендалла, процент попадания в окружность. Для оценки качества классификации и регрессии на основе табличных данных обычно используются - среднеквадратическая ошибка, R2-метрика, матрица ошибок (англ. confision matrix), F1-метрика, AUC ROC, логарифмическая функция потерь [5,7].

После выбора оптимального алгоритма и построения предварительной математической модели следует применить какие-либо способы улучшения качества:

- ансамблирование или гибридные модели (суть этого метода состоит в объединении нескольких математических моделей различных алгоритмов и последовательная или параллельная обработка исходных данных для увеличения точности результатов);
- перекрестная проверка (кросс-валидация, при использовании этого метода для обучения используются все исходные данные, в отличие от традиционного разделения на обучающую и тестовую выборки, плюсы такого подхода особенно заметны на разнородных данных и при их малом объеме);
- подбор параметров по сетке (англ. grid search) и случайный подбор параметров (англ. random search) (этот метод позволяет автоматизировано подобрать параметры алгоритмов, называемые гиперпараметрами, наиболее подходящие для конкретной задачи, отличие состоит в переборе значений в заранее указанном интервале, либо на основе случайно указанных значений);

После того как были подобраны оптимальные параметры (гиперпараметры) алгоритмов нужно сделать возможным использование созданной модели на реальных данных. Для этого есть несколько вариантов:

- создание модулей для какой-либо ГИС (поскольку одним из самых распространенных, общих для ГИС и для библиотек языков программирования стал Python, то внедрение сторонних библиотек в интерфейс настольной ГИС может быть выполнено достаточно просто)
- создание web-обвязки для выполнения скриптов на сервере
- размещение на специализированном облачном сервисе

Заключение

Документы и программный код проведенных исследований в формате программного обеспечения Jupyter Notebook размещен по адресу https://github.com/AlexeyKW/Spatial_ML. Можно сделать выводы о том, что использование современных технологий машинного обучения может привнести новые методы в уже привычные технологические цепочки обработки и использования данных

ДЗЗ. Но для выработок четких алгоритмов и технологий в значимости от решаемой задачи и имеющихся данных потребуются дальнейшие исследования.

Список литературы

1. Колесников А.А., Кикин П.М., Комиссарова Е.В., Грищенко Д.В. Использование машинного обучения для построения картографических изображений // Международная научно-практическая конференция «От карты прошлого – к карте будущего», 28 — 30 ноября 2017, г. Пермь – г. Кудымкар. – с. 110-120
2. Breiman L. Random forests. // Machine learning, T. 45, №. 1, 2001. с. 5–32
3. Badrinarayanan V., Kendall A., Cipolla R. Convolutional encoder-decoder architecture for image segmentation. // IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (12), №. 7803544 , 2017, с. 2481-2495
4. Dai J., He K., Sun J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. // Proceedings of the IEEE International Conference on Computer Vision, 2015, с. 1635–1643.
5. Haug S., Ostermann J. A Crop Weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks. // Computer Vision - ECCV 2014 Workshops. Zurich: Springer, 2014, с. 105–116.
6. Giusti, J. Guzzi, Cires D. C. , He F.-L., Rodrıguez J. P., Fontana F., Faessler M., Forster C., Schmidhuber J., Di Caro G. A machine learning approach to visual perception of forest trails for mobile robots. // IEEE Robotics and Automation Letters, T. 1, №. 2, 2016, с. 661–667
7. Eigen D., Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. // Proceedings of the IEEE International Conference on Computer Vision, 2015, с. 2650–2658
8. Rey, S. J., Smith, R. J. A spatial decomposition of the Gini coefficient. // Letters in Spatial and Resource Sciences, 2013, 6:55–70.
9. Hung C., Nieto J., Taylor Z., Underwood J., Sukkarieh S., Orchard fruit segmentation using multi-spectral feature learning. Intelligent Robots and Systems (IROS) // IEEE/RSJ International Conference on. IEEE, 2013, с. 5314–5320.
10. Mortensen K., Dyrmann M., Karstoft H., Jørgensen R. N., Gislum R. Semantic segmentation of mixed crops using deep convolutional neural network. // International Conference on Agricultural Engineering, 2016.