Инструментальные средства, используемые в рамках решения задачи классификации возрастного интервала при судебно-медицинской идентификации личности

DOI:10.36581/CITP.2020.53.45.009

Н.В. Гридина [1,2], Г.В. Золотенкова[1,2], Ю.И. Пиголкин[1,2] А.И. Рогачев[1]
ФГБУН Центр информационных технологий в проектировании РАН
РФ, г. Одинцово
[2]Кафедра судебной медицины ФГАОУ ВО Первый МГМУ
им. И.М. Сеченова Минздрава России (Сеченовский Университет)
РФ, г. Москва

Аннотация

В данном сообщении дано описание пайплайна, использованного в работе по разработка специализированного комплекса оценки биологического возраста методами машинного обучения. Он состоял из следующих этапов: feature engineering, с двумерной визуализацией данных, анализ информативности признаков с использованием деревьев решений, анализ зависимости качества работы моделей от размерности признакового пространства; сравнительный анализ классификаторов и выбор оптимального алгоритма для решения целевой задачи; построение и анализ матрицы ошибок и ROC-кривой для каждого из классов.

Ключевые слова: алгоритмы классификации, feature engineering, матрица ошибок, ROC-кривая, Random Forest, информативность признаков

В условиях роста числа техногенных катастроф, природных бедствий и террористических актов существенно повышаются требования к качеству, скорости и точности проведения судебно-медицинских экспертиз по отождествлению личности. При массовых катастрофах, как правило, наблюдаются значительные повреждения трупов. В связи с чем, их макроскопическое исследование часто оказывается недостаточным не только для визуального опознания, но и для установления общих идентифицирующих признаков. В подобных случаях используют дополнительные методы исследования, направленные на сбор всей доступной информации о погибшем [1]. Возникает вопрос о способах совокупной оценки сформированного комплекса признаков [2]. В настоящее время большинство методических подходов базируются на использовании множественных уравнений регрессии [1]. Однако, применение линейного регрессионного анализа с позиций современных информационных технологий следует признать утратившими актуальность [3, 4]. Подобные модели не являются масштабируемыми, не обладают достаточной степенью адаптивности и гибкости, а при увеличении количества признаков в исходных данных не всегда позволяют отследить взаимосвязи между параметрами. Одним из наиболее перспективных вариантов решения проблемы является использование методов машинного обучения [5, 6]. В данном сообщении описаны инструментальные средства, которые были

использованы в рамках второго года работы над проектом "Разработка теоретических основ и специализированного комплекса оценки биологического возраста на основе комплексного анализа морфометрических данных в задачах судебной медицины". В рамках работы с данными использовался пайплайн, состоящий из следующих этапов: feature engineering, включающий в себя двумерную визуализацию данных, анализ информативности признаков с использованием деревьев решений, анализ зависимости качества работы моделей от размерности признакового пространства; сравнение различных классификаторов и выбор наиболее подходящего алгоритма для решения целевой задачи; построение и анализ матрицы ошибок и ROC-кривой для каждого из классов. Для решения поставленных задач использовались библиотеки, написанные на языке Python, а так же Jupyter Notebook в качестве интерактивной среды.

Для двумерной визуализации исходных данных, которая помогает оценить наличие или отсутствие кластерной структуры в данных, использовались библиотеки uMAP и класс TSNE из модуля manifold библиотеки scikit-learn. Метод uMAP является наиболее свежим примером метода для нелинейного снижения размерности пространства. Далее создается граф в низкоразмерном пространстве и приближается к исходному. В качестве минимизируемого функционала выступает дивергенця Кульбака Лейблера. При построении всех графиков использовалась библиотека matplotlib. Для отбора признаков использовался метод RFECV (sklearn), позволяющий определить информативность признаков путем перебора подмножеств признаков с последовательным удалением наименее информативных из рассматриваемого в данных момент множества. Для улучшения качества работы данного подхода использовалась модификация, позволяющая проделать эксперименты на разных фолдах. Для выбора наиболее оптимального размера признакового пространства проводились эксперименты, в ходе которых анализировалась зависимость качества работы алгоритмов, а именно - значения метрики f1-score, от размера признакового пространства. Реализация алгоритмов бралась из библиотеки sklearn и catboost. После выбора наиболее информативных признаков и размерности признакового пространства на предыдущем этапе, проводилось сравнение наиболее популярных алгоритмов классификации, используемых в машинном обучении. Поскольку часто наблюдался дисбаланс классов в исходных данных, как и на предыдущем этапе, использовалась метрика f1-score, являющаяся наиболее подходящей в такой ситуации. Для каждой модели производилась настройка гиперпараметров с применением класса Grid Search CV библиотеки sklearn, реализующего перебор параметров по сетке. При составлении сеток учитывался факт возможности переобучения, поэтому в качестве перебираемых значений указывались лишь те, которые позволяли бы ограничивать склонность алгоритмов к переобучению. Так, например, для деревьев решений использовалась небольшая максимальная глубина, иными словами, применялся pruning и регуляризация. Эксперименты так же проводились на 5 фолдах. Сравнивались модели, которые получали

наибольший f1-score. Наилучшая модель использовалась для дальнейших экспериментов. Для анализа классификации каждого из классов строилась матрица ошибок, позволяющая понять, как именно "ошибается" модель, куда относятся неправильно классифицированные объекты, выделить классы, работа с которыми дает наибольшее количество ошибок. Полученная матрица визуализировалась в виде тепловой карты. Благодаря данной информации в ходе экспериментов варьировались возрастные интервалы каждого класса и количество классов с целью поиска оптимального соотношения качества работы классификаторов и внутриклассовой возрастной дисперсии. Производился анализ ROC-кривых, построенных для каждого из классов. Для каждого из них строилась отдельная модель, задача которой была отделять каждый конкретных класс от всех остальных данных. Далее, по полученным результатам строилась кривая, форма графика и площадь которой позволяли выявить наиболее проблемные классы. Использовалось как самописное решение, так и обновленная в процессе работы над проектом библиотека sklearn-plot.

Заключение

Описанный пайплайн использовался при работе со всеми данными в рамках проекта и может быть применен для реализации задач, с использованием для обучения исходных данных малого объема. В подобных случаях крайне важен процесс отбора признаков и контроль за отсутствием переобучения моделей. При этом использование библиотек, реализованных на языке Python не вызывает трудностей из-за времени работы, так как нет потребности в построении сложных моделей, обучение которых является ресурсоемким процессом, требующим использование GPU. Более того, реализации некоторых алгоритмов, таких как Catboost или Random Forest, поддерживают использование распараллеливания задач.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта 19-07-00982 а

Литература

1. *Пиголкин Ю.И., Золотенкова Г.В., Березовский Д.П.* Методологические основы определения возраста человека. // Судебно-медицинская экспертиза. – 2020. – Т. 63. – №2. – С. 58-63.

2. *Золотенкова Г.В., Гридина Н.В., Солодовников В.И., Труфанов М.И., Пиголкин Ю.И.* Вычисление биологического возраста индивидуума с использованием новейших информационных технологий и построение перспективного интеллектуального программно-аппаратного комплекса. // Судебно-медицинская экспертиза. – 2019. – Т. 62. – №3. – С. 42-47.

3. *Kimmerle E. H, Jantz R. L., Konigsberg L. W., Baraybar J. P.* Skeletal estimation and identification in American and East European populations // Journal of Forensic Sciences. – 2008. - Vol. 53, № 3. – P. 524–532.

4. *Ferrante L., Skrami E., Gesuita R., Cameriere R.* Bayesian calibration for forensic age estimation // Statistics in Medicine. – 2015. – Vol. 34. – № 10. – P. 1779–1790.

5. *Гридина Н.В., Золотенкова Г.В., Рогачев А.И.* Использование классификаторов для целей судебно-медицинской идентификации личности (диагностики возраста). // Биомедицинская радиоэлектроника. – 2019. – Том 22 – №5. – С. 38-44.

6. *Золотенкова Г.В., Гридина Н.В., Солодовников В.И.* Алгоритм вычисления биологического возраста индивидуума с использованием новейших информационных технологий. // В сб.: Информационные технологии и математическое моделирование систем – 2018 – С. 151-154.