УДК 004.896

ЭВОЛЮЦИОННОЕ МОДЕЛИРОВАНИЕ СТРАТЕГИЙ ОБУЧЕНИЯ ДЛЯ МОДЕЛЕЙ ОСНОВАННЫХ НА АДДИТИВНОЙ РЕГУЛЯРИЗАЦИИ

Ходорченко М.А. (Университет ИТМО)
Научный руководитель – к.т.н., доцент ИДУ Бутаков Н. А.
(Университет ИТМО)

Тематическое моделирование является популярным методом обработки текста, который обеспечивает интерпретируемое представление документа. BigARTM - хорошо известная библиотека, позволяющая работать с текстами из разных предметных областей, а также с разными статистическими особенностями. Однако, из-за ее гибкости и большого количества регуляризаторов, становится сложно найти оптимальную стратегию обучения модели для создания качественных интерпретируемых тематик. В работе предлагается подход, который формализует проблему в виде вектора параметров и успешно его оптимизирует с помощью генетического алгоритма. Экспериментальное исследование проведенное на английских и русских наборах данных показывает, что предлагаемый метод эволюционной оптимизации может улучшить качество получаемых тематических моделей и превзойти байесовскую оптимизацию как по скорости поиска, так и по качеству результата.

В настоящее время тематическое моделирование - наиболее активно используемый метод обработки текста без учителя. В свою очередь, подход, основанный на аддитивной регуляризации, является одним из самых высокоуровневых разновидностей моделирования благодаря гибкой системе регуляризаторов, которая помогает эффективно выделять темы, скрытые в документе учитывая особенности входных данных (длина текста, размер словаря и т. д.). Однако, для определения оптимальных значений для регуляризаторов, числа тем и количества итераций обучения, исследователям приходится подготавливать набор моделей с разными параметрами, затем вручную проверять полученные результаты и, руководствуясь внутренними эвристиками и здравым смыслом, дорабатывать модель. Автоматизация процесса выбора наиболее оптимальной по критерию качества модели в условиях фиксированного количества возможных оцениваний функционала качества позволит ускорить процесс работы с текстом, улучшить качество получаемых векторов тем, а также получить более предметно-ориентированные темы при увеличении их количества.

Для оптимизации модели BigARTM было предложено создание вектора, состоящего из процесса инициализации и изменения параметров регуляризаторов, количества фоновых тем, а также количества итераций обучения между установкой и сменой значений. В качестве метрики качества была выбрана когерентность 50 наиболее вероятных слов в теме, которая наиболее соотносится с оценкой качества, которую присваивает теме человек. Несмотря на то, что по нашим экспериментам на 400 тем примерно в 40% случаях наиболее связные темы, отмеченные асессорами, не являлись таковыми по оценке когерентности, все же наблюдается явная связь между высоким значением оценки и интерпретируемостью темы, в отличие от других используемых метрик качества, таких как чистота и контрастность ядра, а также оценки, основанной на w2v эмбеддингах.

Регуляризаторы и функция качества были подобраны с целью выполнения нескольких требований к тематической модели:

- **разнообразие тем** обеспечивается введением двух отдельных декоррелляторов на матрицы распределений слов по темам для предметных и фоновых тематик;
- **понятность темы для человека** регулируется выбранной метрикой качества модели;
- **низкое количество фоновых слов в предметных темах** обеспечивается возможностью задания некоторого количества фоновых тем, которые определяются заданием регуляризаторов сглаживания.

Для проверки качества работы автоматической оптимизации было произведено сравнение нескольких алгоритмов оптимизации, а именно, поиска по сетке, байесовской оптимизации, дифференциальной эволюции и модифицированного генетического алгоритма.

Сравнение проводилось на пяти наборах данных, из которых два - на русском языке, три - на английском.

По полученным результатам экспериментального исследования можно сделать вывод о том, что настроенный эволюционный алгоритм работает оптимальнее в терминах скорости и качества, т. к. позволяет получить до 26% более оптимальные результаты метрики качества, по сравнению с байесовской оптимизацией на четырех наборах данных из пяти.

Ходорченко М. А. (автор)

Бутаков Н. А. (научный руководитель)