УДК 000.00

МЕТОДЫ ОТБОРА ПРИЗНАКОВ ДЛЯ МЕДИЦИНСКИХ ДАННЫХ В ТАБЛИЧНОЙ ФОРМЕ

Балабаева К.Ю. (Университет ИТМО),
Научный руководитель – **к.т.н., доцент Ковальчук С. В.**
(Университет ИТМО**)**

Доклад посвящен проблеме сокращения пространства признаков для более качественной работы алгоритмов машинного обучения и снижения риска их переобучения. В работе будет представлен обзор методов для отбора признаков их сравнительная оценка для задачи мульти-классификации, а также предложение альтернативного метода отбора признаков.

Введение. Одним из наиболее популярных типов задач в области машинного обучения является обучение с учителем, которое подрывает использование набора данных с соответствующими метками для каждого экземпляра. Такими задачами являются, например, регрессия, бинарная и мульти-классификация в зависимости от типа целевой переменной.
При обучении с учителем входные данные обычно состоят из матрицы признаков и целевого вектора. Производительность алгоритмов машинного обучения сильно зависит от количества и вариативности выборок, представленных в обучающем наборе данных. Однако увеличение числа переменных может привести к проклятию размерности. Эта проблема относится к более высокому риску переобучения, особенно если количество признаков превышает количество наблюдений. Другая проблема, возникающая из-за большой размерности, заключается в том, что наблюдения в пространстве высокой размерности становятся равноудаленными, что затрудняет их группировку или классификацию. Чтобы решить эту проблему, мы должны уменьшить пространство признаков. Есть несколько способов справиться с этим: извлечение признаков (PCA, LDA, Transformer) или выбор признаков, который будет обсуждаться более подробно в дальнейших разделах. В целом, целью всех методов является сокращение количества столбцов в обучающем наборе данных.
По сравнению с извлечением признаков методы выбора признаков более прозрачны и объяснимы, поскольку сжатое признаковое пространство состоит из исходных переменных в данных. Более того, такое сокращение может снизить время обучения и способствовать повышению точности модели.
Основная часть. В работе мы сравниваем несколько методов отбора признаков на случай прогнозирования стадии хронической сердечной недостаточности. Мы также представляем подход байесовского вывода к задаче выбора признаков. В качестве критериев мы оцениваем алгоритмы отбора с использованием f-score с макро-усреднением в качестве показателя эффективности, стабильность модели, использование k-fold кросс-валидации и интерпретируемость процесса отбора признаков.
Выводы. Мы предложили основанный на фильтрации подход для решения задачи отбора признаков, основанный на байесовском выводе и вероятностном моделировании. Этот метод продемонстрировал прирост к точности модели, а также устойчивость к проблеме переобучения. В наших экспериментах мы сравнили этот метод с другими алгоритмами выбора признаков и представили результаты, касающиеся их стабильности, точности и объяснимости.

Балабаева К.Ю. (автор) Подпись

Ковальчук С.В. (научный руководитель) Подпись