# Метод генерации признаковых описаний, основанный на расстояниях до эталонов, в биомедицинских исследованиях

***Сенько Олег Валентинович***[1,2] senkoov@mail.ru
***Салманов Махир Юсиф оглы***[1]⋆ sy.mahir@gmail.com
***Брусов Олег Сергеевич***[3] oleg.brusow@yandex.ru
***Матвеев Иван Алексеевич***[2] matveev@ccas.ru
***Кузнецова Анна Викторовна***[4] azfor@yandex.ru

[1]Москва, МГУ им. М.В.Ломоносова
[2]Москва, ФИЦ ИУ РАН
[3]Москва, ФБГНУ НЦПЗ РАН
[4]Москва, Институт биохимической физики им. Н.М.Эмануэля

Существующие методы машинного обучения во многом основываются на принципе компактности, предполагающие близкое расположение описаний объектов со сходными значениями целевой переменной $Y$. Среди различных способов реализации принципа компактности может быть выделен подход, основанный на использовании метрик заданных на множестве признаковых описаний. В качестве примера можно привести метод k-ближайших соседей и метод опорных векторов. В первом случае прогноз $Y$ в точке $\boldsymbol{x}$ вычисляется по значениям $Y$ на $k$ ближайших к точке $\boldsymbol{x}$ объектов обучающей выборки. В наиболее популярном варианте метода опорных векторов прогноз значения $Y$ вычисляется по линейной комбинации ядерных функций, каждая из которых зависит от расстояния между $\boldsymbol{x}$ и одним из опорных объектов обучающей выборки. Высокая эффективность методов машинного обучения, основанных на использовании при решении задач распознавания расстояний до эталонных объектов, подтверждена многочисленными экспериментами. Следует также отметить, что методы, основанные на оценке общей похожести рассматриваемого объекта с ранее встречавшимися случаями, часто используются при принятии решений, например, в медицине. Способы применения расстояний до эталонных объектов не могут сводится только лишь к использованию линейных комбинаций монотонно трансформированных расстояний. Альтернативным подходом является использование набора расстояний между произвольным объектом $\boldsymbol{x}$ и эталонными объектами в качестве нового векторно-

го описания $\boldsymbol{z(x)}$, по которому далее производится прогноз значения $Y$ с помощью алгоритма, который был обучен на выборке, состоящей из полученных таким же образом новых векторных описаний объектов исходной обучающей выборки. При этом обучение может производиться с помощью самых разных технологий, включая случайный лес и градиентный бустинг. Использование ансамблей решающих деревьев позволяет не только описывать сложные нелинейные зависимости, но и добиваться присущей ансамблевым методам более высокой устойчивости обучения.

Указанный подход был использован для решения задачи диагностики шизофрении по характеру тромбодинамики, то есть по динамике образования спонтанных фибриновых сгустков. Необходимость использования рассматриваемой технологии при решении именно этой задачи связана с тем, что диагностику предполагается производить по кривым, показывающей временную зависимость интенсивности отражённого света на фибриновых сгустках.

Среди существующих подходов работы с такими кривыми можно выделить способ, основанный на подсчёте для каждой кривой нескольких характеризующих её параметров. Недостатком данного подхода является определённая утрата информации. Альтернативным способом является использование признаков, соответствующих отдельным временным точкам отсчёта. Каждый из таких признаков принимает значения интенсивности отражённого света соответствующей точке. Недостатками подхода являются высокая размерность при большом числе точек отсчёта и высокая коррелированность признаков. В наших исследованиях сравнивалась эффективность методов, основанного на расстояниях до эталонных объектов, и метода, основанного на использования в качестве признаков измерений на всевозможных точках отсчёта, при распознавании группу из пациентов с шизофренией и контрольной группы из испытуемых. В рамках подхода, основанного на расстояниях до эталонов исследовалась эффективность использования трёх метрик: эвклидовой метрики, косинусной метрики и метрики Минковского (p=1). Использовались два способа отбора эталонных объектов:

1. отбирались опорные вектора из всех объектов на основе метода SVC (Support Vector Classifier);
2. в качестве эталонов использовались все объекты обучающей выборки.

Метод кросс-валидации с тестированием на каждом шаге на одном объекте (Leave One Out) был использован для сравнения эффективности логистической егрессии (ЛР), случайного решающего леса (СРЛ) и градиентного бустинга (ГБ) над решающими деревьями. Оценка доверительного интервала результата классификации ROC AUC на основе бутстрапа показала, что полученные результаты стабильны. Для обеспечения наглядности результатов и возможности их интерпретации наряду с многофакторными методами распознавания использовался также метод интеллектуального анализа данных, основанный на оптимальных достоверных разбиениях признакового пространства (метод ОДР), в рамках подхода, представленного в работе [1].

Значения ROC AUC для показавшей наилучшие результаты метрики Минковского (p=1) представлены в таблице. В качестве эталонов использовались все объекты обучающей выборки и объекты, отобранных SVC:

| | ЛР | СЛ | ГБ |
|---|---|---|---|
| с отбором объектов | 0.737 | 0.778 | 0.706 |
| без отбора | 0.723 | 0.778 | 0.762 |

Проведённые исследования выявили существование выраженной внутренней структуры в данных, генерируемых с использованием расстояний до эталонных объектов. Внутренняя структура при этом характеризует также различиями между группой пациентов с шизофренией и контрольной группой.

Работа выполнена при поддержке РФФИ, проект 20-01-00609.

[1] *Доровских И.В., Сенько О.В., Чучупал В.Я., Докукин А.А., Кузнецова А.В.* Исследование возможности диагностики деменции по сигналам ЭЭГ с помощью методов машинного обучения. // Математическая биология и биоинформатика, 2019. т.14, вып.2, С. 543–553.