Оценка оптимального объёма выборки для задач классификации*

***Харатян А. С., Катруца А. М.*[1], *Стрижов В. В.*[2]**
haratyan.as@phystech.edu; aleksandr.katrutsa@phystech.edu;
strijov@phystech.edu
[1]Московский физико-технический институт, Москва, Россия;[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, Москва, Россия

В статье рассматривается задача выбора оптимального числа объектов выборки для их классификации. Исследуется использование порождающих и разделяющих вероятностных моделей бинарной классификации. Обсуждается проблема медицинской диагностики пациентов. Определяется понятие достаточности объёма выборки. Показывается, какими методами возможно выбрать оптимальное количество объектов, обеспечивающее необходимую точность классификации объектов . В работе рассматривается, применение каких критериев выявляет наилучшее качество классификации. Приводится теоретическое и практическое обоснование предложенных критериев. Используется модель логистической регрессии.

Ключевые слова: *определение оптимального объёма выборки, логистическая регрессия, кросс-валидация.*

1 Введение

Работа посвящена оценке оптимального объёма исследуемой выборки применительно к проблемам медицинской диагностики. Рассматриваются биомедицинские данные пациентов как выборка. Каждый пациент набором признаков. Получение данных о пациентах требует немалых средст. В случае, если количество данных избыточно, то их измерения приносят крайне неоправданные расходы. В связи с этим поднимается вопрос оптимального количества измерений. Ввиду дороговизны анализов всех признаков оценка измерений должна быть точной.

Для нахождения оценки используется модель логистической регрессии [1]. Стандартной практикой является использование статистических методов [2] для оценивания объёма данных при помощи логистической регрессии. Введём понятие устойчивости модели в отношении объёма выборки. Будем называть модель устойчивой, если при изменении малом изменении объёма параметры модели меняются незначительно. Если размер выборки крайне мал и недостаточен, то параметры модели меняются скачкообразно при увеличении объёма.Соответственно, увеличивая объём выборки, мы повышаем устойчивость модели. В качестве показателя устойчивости для моделей будем использовать расстояние Кульбака-Лейблера. Для того чтобы показать отличие моделей в устойчивости будем использовать разность усредненных значений расстояний К-Л, вычисленных на разных выборках одного и того же объёма. Если объекты порождены одинаковым распределением, то при росте объёма выборок разность расстояний К-Л между моделями падает. Достигнув необходимого показателя устойчивости, можно легко вычислить оптимальный размер данных.

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Консультант: Катруца А. М.

Для вычислительного эксперимента используются реальные данные 569 пациентов с 30 признаками и метками об опухоли молочной железы: доброкачественная или злокачественная. Как описано выше, будем обучать модели на разных подвыборках и после достижения необходимого уровня устойчивости получим оптимальный объём данных.

Литература

[1] *Hosmer Jr David W, Lemeshow Stanley, Sturdivant Rodney X.* Applied logistic regression. — 2013. Vol. 398.

[2] *Demidenko Eugene.* Sample size determination for logistic regression revisited // Statistics in medicine, 2007. Vol. 26. No. 18. P. 3385–3397.

Received