

Раннее прогнозирование достаточного объема выборки для обобщенно линейной модели

Жолобов В. А. Малиновский Г. Стрижов В. В.

Исследуется проблема планирования эксперимента. Задача раннего прогнозирования важна в медицинском применении, особенно в случаях дорогостоящих измерений иммунных биомаркеров. Решается задача оценивания достаточного объема выборки по данным. Предполагается, что выборка является простой. Она описывается адекватной моделью. Иначе, выборка порождается фиксированной вероятностной моделью из известного класса моделей. Объем выборки считается достаточным, если модель восстанавливается с достаточной достоверностью. Исследуется зависимость функции ошибки от объема данных. Исследуется зависимость модели от редуцированной матрицы ковариации параметров *GLM*. Требуется, зная модель, оценить достаточный объем выборки на ранних этапах сбора данных. Созданы алгоритмы определения достаточного объема данных на раннем этапе. Проведен вычислительный эксперимент с использованием синтетических данных.

1 Введение

Работа посвящена задаче оценивания достаточного объема выборки на раннем этапе сбора данных. Задача возникла из условия, когда необходимо провести крупное исследование, а сбор данных является дорогостоящим занятием. Для примера можно взять медицинское исследование, такой как анализ крови. Существуют такие виды анализа крови, которые стоят достаточно приличных денег для людей. Для того, чтобы снизить стоимость данных для исследований в несколько раз необходимо построить модель, а для модели нужно собрать выборку. Поэтому в данной работе рассматривается задача построения алгоритма для предсказания оптимального набора данных при заданной модели. Предлагаемый в данной работе метод должен на малой выборке спрогнозировать ошибку на пополняемой большой. Выборка считается простой, то есть удовлетворяет простому распределению. Предлагается использовать два разных метода: полного перебора и генетический алгоритм.

Кроме этих методов ранее задача прогнозирования достаточного объема выборки решалась в работе [1]. Здесь был предложен метод, основанный на технике кросс-валидации и расстоянии Кульбака-Лейблера между двумя распределениями параметров модели, оцениваемых на аналогичных подмножествах данных. Похожая задача информационного поиска решалась в работах [2,3]. Здесь для создания простых структурированных функций информационного поиска используется модернизированный генетический алгоритм. Эвристика генетического алгоритма заключается в том, что он способен работать при стагнации признаков.

В данной работе используются два метода. Основной из них — это метод полного перебора. Необходимо подобрать такую функцию, которая является монотонной и достаточно гладкой, то есть гарантируется непрерывная дифференцируемость до второго порядка. Метод заключается в том, что он аппроксимирует зависимость функции ошибки от объема данных по малому объему выборки, чтобы с достаточной точностью предсказывать ее поведение. Считается, что модель в этой задаче задана и зависит от редуцированной матрицы ковариации параметров *GLM*. Также предложен способ генерации такой функции через генетический алгоритм.

Вычислительный эксперимент проводится на синтетических данных *Boston Housing* и *Diabets*. Вначале реализуем метод полного перебора. Разделяем выборку на два множества. Строим два графика поверхности выборок: первую получаем с помощью бутстрепа [4] для подвыборки фиксированного объема, вторую через аппроксимацию. Чтобы получить аппроксимирующую поверхность, решается оптимизационная задача. Затем повторяем действия, используя уже для поиска аппроксимирующей функции генетический алгоритм. Решение этой задачи позволит находить оптимальное значение объема выборки.

2 Постановка задачи

Задана выборка $\mathfrak{D} = [\mathbf{X}, \mathbf{y}] = \{(\mathbf{x}_i, y_i)\}$, являющаяся *i.i.d.* произвольных значений, сгенерированных неизвестным распределением. Обозначим объем выборки как $m = |\mathfrak{D}|$. Решается задача раннего прогнозирования объема минимально необходимой выборки. Под ранним прогнозированием понимается прогнозирование при единственно заданном объеме m_0 объема m^* , необходимого для построения адекватной модели. При решении задачи поиска оптимального объема выборки считаем заданной модель, то есть функцию $f: \mathbf{x} \mapsto y$. Для отображения f задаются функции ошибки, указывающая на точность модели в случаях линейной и логистической регрессии

$$S = S(\mathfrak{D}, f, w^*, m) = \|f - y\|^2$$

$$S(\mathbf{w}, \mathfrak{D}) = \sum_{i=1}^n \log(1 + \exp(-y_i(\mathbf{w}, \mathbf{x}_i)))$$

Здесь предполагается фиксированность всех параметров, кроме объема данных. Рассматривается несколько постановок задач обобщенной линейной модели. Первой будет задача линейной регрессии. Пусть дано множество из m пар (x_i, y_i) , $i = 1, \dots, m$, а также пусть назначена линейная модель с аддитивной случайной величиной ε_i

$$y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon_i$$

Требуется найти параметры регрессионной модели \mathbf{w}^*

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^n} S(\mathbf{w} | \mathfrak{D}),$$

где S — функция ошибки.

Постановка задачи логистической регрессии. Считаем, что здесь принята модель логистической регрессии, согласно которой свободные переменные x и зависимая переменная y связаны зависимостью

$$z = w_0 + \sum_{j=1}^n w_j x_j.$$

$$y = \frac{1}{1 + \exp(-z)} + \varepsilon,$$

Обозначим $p_i = f(w, x_i)$. Функция ошибки

$$S(\mathbf{w}, \mathfrak{D}) = \sum_{i=1}^n \log(1 + \exp(-y_i(\mathbf{w}, \mathbf{x}_i)))$$

В задаче поиска оптимального состава признаков $X_{\mathcal{A}}$ требуется оптимальный набор признаков $n^* = |A|$ и

$$A^* = \arg \min Q(A|\mathbf{w}, \mathfrak{D})$$

Обозначим зависимость значения функции ошибки от выборки как $S(m) = S(m|\mathbf{w}, \mathcal{A}, \mathfrak{D})$. Так как функция S зависит от объема данных, то и при каждом значении m решается отдельная обобщенно-линейная задача $w^* = w^*(m)$.

В задаче раннего прогнозирования предполагается заданной ожидаемая точность r . По заданной выборке $\mathfrak{D}(m_0)$ необходимо построить функцию $\varphi(m)$, аппроксимирующую функцию $S(m)$. Подбор функции $\varphi : \mathbb{N} \rightarrow \mathbb{R}_+$ осуществляется из заданного класса с ограничением, связанным с принадлежностью функции $\varphi(m)$ классу дважды непрерывно дифференцируемых функций $E \subset \mathbb{C}^2(\mathbb{R}_+)$. Для поиска функции решается условная оптимизационная задача

$$\varphi^*(m) = \arg \min_{\varphi(m) \in E, \varphi'(m) \geq 0, m \in [0, m_0]} \|\varphi(m) - S(m)\|_1$$

Для решения задачи выбирается функция $\varphi^*(m)$ и ищется прообраз в значении заданной точности

$$m^* = (\varphi^*)^{-1}(r)$$

Литература

- [1] Anastasiya Motrenko, Vadim Strijov, and Gerhard-Wilhelm Weber. Sample size determination for logistic regression. *Journal of computational and applied mathematics*, ISSN 0377-0427, 255(1):743–752, 2014.
- [2] A. S. Kulunchakov and Vadim V. Strijov. Generation of simple structured information retrieval functions by genetic algorithm without stagnation. *Expert Syst. Appl*, 85:221–230, 2017.
- [3] Parantapa Goswami, Simon Moura, Eric Gaussier, Massih-Reza Amini, and Francis Maes. Exploring the space of ir functions. pages 372–384, April 14 2014.
- [4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Поступила в редакцию