Оценка оптимального объема выборки для исследований в медицине*

Сеппар[1] ***А. Д. Михеев***[1] ***М. С Макаренко***[1] ***С. Д Мурлатов***[1] ***С. Ю Коноплёв***[1] ***М. Д Катруца***[1] ***А. М. Стрижов***[2] ***В. В.***

[1]Московский физико-технический институт

[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Данная задача посвящена оценка оптимального объема выборки для исследований в медицине. В условиях недостаточного числа дорогостоящих измерений требуется спрогнозировать оптимальный объем пополняемой выборки. В качетсве алгоритма используется серия эмпирических алгоритмов оценки объема выборки.

Ключевые слова: *оптимальный объем выборки*, *расстояние Кульбака-Лейблера*.

Введение

Цель исследования: Целью работы является создание и теоретическое обоснование методов оценки объема многофакторных выборок, учитывающих вид модели классификации и более точных по сравнению с известными методами; создание методов классификации малых выборок.

Предмет исследования: Оценить минимальный объём выборки — количество производимых измерений некоторого параметра или набора параметров, необходимый для выполнения некоторых ранее сформулированных условий.

Исследуемая проблема: Исследование направлено на решение проблемы выбора моделей при классификации выборок малой мощности. По заданной выборке, включающей многокритериальное описание объектов и метки класса объектов, требуется получить оценку структурных параметров, получить асимптотическую оценку необходимого объема выборки и указать предпочтительный подход к решению задачи классификации. Для классификации объекта требуется получить оценку параметров выбранной модели и выполнить анализ ошибок классификации.

Решаемая в данной работе задача: В данной работе основное внимание уделяется байесовским методам оценки объёма выборки. Оценка объёма выборки в байесовской постановке включает оценку апостериорного распределения $\mathsf{p}(D|w)$ параметров модели. При отсутствии наблюдаемых данных, апостериорное распределение

$$p(w, D) = \frac{p(D|w)p(w)}{p(D)}$$

совпадает с априорным $\mathsf{p}(w)$, как в классических методах оценки объёма выборки. Разница между вторым и третьим случаями заключается только в способе оценки распределения $\mathsf{p}(w|D)$ — на основе сэмплированных, либо реально наблюдаемых данных.

Предлагаемое решение: Статистические методы позволяют оценить объем выборки, исходя из предположений о распределении данных и информации о соответствии наблюдаемых величин предположениям нулевой гипотезы. В случае, если объем исследуемой

Научный руководитель: Стрижов В. В. Авторы: А.М. Катруца, В.В. Стрижов В В. Координатор: Т. Гадаев

выборки достаточен или избыточен, возможно применение методов, основанных на наблюдении за изменением некоторой характеристики процедуры построения модели при увеличении объема выборки. В частности, наблюдая за отношением качества прогнозирования на контрольной выборке и обучающей выборке, определим достаточный объем выборки как соответствующий началу переобучения. Таким же образом производится оценка объема выборки в рамках предлагаемого метода: предлагается считать объем выборки достаточным, если расстояние между распределениями, оцененными на подвыборках данного объема, достаточно мало. Такой подход не требует дополнительного обобщения на случай многих переменных. Кроме того, оценку можно производить как при наличии предположений о распределении данных, так и в их отсутствие.

Анализ сильных и слабых сторон предлагаемого решения: Недостатком данного подхода является то что количественные оценки возможно получить лишь в случае, когда объем выборки избыточен. В противном случае метод позволяет лишь определить, является ли текущий объем выборки достаточным

Постановка задачи

Постановка задачи классификации: Рассмотрим выборку вида $D_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, где $\mathbf{x}_i \in \mathbb{R}^n$ — описание i-го элемента выборки, а $y_i \in \mathcal{Y}$ — его метка класса. Введем обозначение $D_m = (X, \mathbf{y})$, где $\mathbf{y} = [y_1, \ldots, y_m]^R$, $= [\mathbf{x}_1^R, \ldots, \mathbf{x}_m^R]^R$ — матрица плана.

В задаче классификации необходимо подстроить отображение наблюдаемых данных $\in \mathbb{R}^m$ на множество меток класса $\mathcal{Y}$. Будем называть функцию $a : \mathbb{R}^m \to \mathcal{Y}$ *классификатором*. В данной работе рассматривается подход к построению классификатора, основанными на максимизации правдоподобия данных.

В данной работе предполагается, если не указано иное, что выборка (y, X) простая, то есть объекты (y_i, x_i), $i = 1, \ldots, m$ — случайные, независимые величины из одного распределения. Будем называть *гипотезой порождения данных* совокупность предположений о виде неизвестного распределения (y, X) элементов выборки D_m. В данной работе основное внимание уделяется случаю, когда предположения об истинном распределении данных формулируется с точностью до некоторого неизвестного параметра θ. Предполагается, что функция плотности распределения $p(y|X)$ принадлежит параметрическому множеству функций

$$\mathcal{F} = \{f(y, X, \theta), \theta \in \Theta\},$$

то есть найдется такое значение параметра $\theta^* \in \Theta$, что

$$\mathcal{P}(y, X) = f(y, X, \theta^*).$$

В данных обозначениях предположения о независимости и однородности выборки означают, что каждая функция $f \in$ представима в виде произведения

$$f(y, X, \theta) = \prod_{i=1}^{m} g(y_i, x_i, \theta)$$

по элементам выборки.

Пусть фиксировано некоторое значение параметра $\theta \in \Theta$. Тогда классификатор $a(x)$ имеет вид

$$a(x) = \arg\max_{y \in \mathcal{Y}} (y, x, \theta).$$

Таким образом, для решения задачи классификации в данной постановке необходимо оценить значение параметра θ, наиболее согласованного с наблюдаемыми данными.

Постановка задачи определения оптимального объёма выборки: При заведомо недостаточном количестве m объектов выборки D с распределением $\mathsf{P}(\boldsymbol{x}, y)$ для обучения устойчивой логистической регрессии, требуется найти такое количество m^* объектов выборки, которое будет достаточным и, кроме того никакая другая выборка с количеством объектов M^* большей m^* не приносит новой информации о распределении параметровом логистической регрессии. Пусть $\mathbb{P}(\boldsymbol{w}|m)$ суть распределение весов классификатора, который был обучен на выборке D_m с распределением $\mathbb{P}(\boldsymbol{x}, y)$. Тогда имеем

$$m^* = \min m \in \mathbb{N} \; \forall k \in \mathbb{N} : \rho(\mathsf{P}(\boldsymbol{w}|m), \mathsf{P}(\boldsymbol{w}|m+k)) < \varepsilon$$

где ρ — некоторая функция расстояния между распределениями, а ε — заданный порог, в качестве задачи прогнозирования оптимального объёма выборки.

Базовый метод

Предлагается использовать расстояние Кульбака-Лейблера для определения расстояния между распределениями. Будем считать распределение векторов параметров нормальным ввиду простоты выборки, а также учитывая факт подчинения целевой переменной Бернулливскому распределению. Найдем набор оптимальных векторов параметров, обучая логистическую регрессию на подвыборках размеров $m - l$ исходной выборки m. Для двух подвыборок размера m и $m + k$ из нашего набора имеем

$$\rho(\mathsf{P}(\boldsymbol{w}|m), \mathsf{P}(\boldsymbol{w}|m+k)) = KL(\mathcal{N}(\hat{\boldsymbol{w}}_1, \boldsymbol{A}_1)||\mathcal{N}(\hat{\boldsymbol{w}}_2, \boldsymbol{A}_2)) =$$

$$\frac{1}{2}\left(tr(\boldsymbol{A}_2^{-1}\boldsymbol{A}_1) + (\hat{\boldsymbol{w}}_1 - \hat{\boldsymbol{w}}_2)^T\boldsymbol{A}_2^{-1}(\hat{\boldsymbol{w}}_1 - \hat{\boldsymbol{w}}_2) - n + \ln\left(\frac{\det(\boldsymbol{A}_2)}{\det(\boldsymbol{A}_1)}\right)\right),$$

где $\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2$ есть матожидания, а $\boldsymbol{A}_1, \boldsymbol{A}_2$ - матрицы ковариации векторов параметров их нормального распределения.

Пусть m^* оптимальный объём выборки, по которой распределение $\mathsf{P}(y|\boldsymbol{x})$ восстанавливается точно. Малое расстояние

$$\rho(\mathsf{P}(\boldsymbol{w}|m-l), \mathsf{P}(\boldsymbol{w}|m^*))$$

говорит о близости выборки $m-l$ к оптимальному размеру выборки. График зависимости расстояния Кульбака-Лейблера от m достигает своего насыщения при $m = m^*$, ввиду малости изменения

$$\rho(\mathsf{P}(\boldsymbol{w}|m^*+l), \mathsf{P}(\boldsymbol{w}|m^*))$$

при устремлении l к бесконечности. Получим данную зависимость следующим образом. Меняя размер выборки с шагом равным 1, получим l распределений $\mathsf{P}(\boldsymbol{w}|m-l)$, а значит и искомую зависимость. Таким образом найдем m^* оптимальное как точку начала насыщения графика.

Вычислительный эксперимент

Реальные данные: Проведем исследование на наборе данных о сердечно-сосудистых заболеваний. Рассматривается 5 признаков на 270 объектах. Рассматриваемая модель -

логистическая регрессия. На разных размерах выборки отслеживаются следующие глобальные параметры распределения:

1) логарифмические потери
2) расстояния Кульбака-Лейблера
3) след матрицы ковариации
4) Евклидово расстояние средних

При проекции средних векторов параметров на двумерное пространство наблюдаемое распределение аппроксимируется нормальным.

Синтетические данные: Проведем эксперимент на вручную сгенерированых данных с нормальным распределением так, чтобы на них можно было протестировать работу алгоритмов классификации. Размер синтетических данных соответсвует размеру исходных данных. Проведя эксперимент, было выяснено, что рассматриваемые выше глобальные параметры в модели на реальных данных стабилизируются быстрее, а именно при размерах выборки около 65 объектов.

Вывод

Наибольшую ценность из вышеперичисленных параметров составила след матрицы ковариации. Данный компонет выходит на плато при размерах выборок более 80 объектов.

Литература

[1] Stephen Bush. Sample size determination for logistic regression: A simulation study. *Communications in Statistics - Simulation and Computation*, 44(2):360–373, 2015.

[2] Michael P. Cohen. Sample size determination. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1269–1271. Springer, 2011.

[3] Sample Size Determination. Ralph b. dell, steve holleran, and rajasekhar ramakrishnan. *Ilar Journal*, 2002.

[4] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1), feb 2012.

[5] Ira Cohen Forman George. Learning from little: Comparison of classifiers given little training. In *ECML/PKDD*, 2004.

[6] LAWRENCE JOSEPH, ROXANE DU BERGER, and PATRICK BE LISLE. Bayesian and mixed bayesian/likelihood criteria for sample size determination. *STATISTICS IN MEDICINE*, 1997.

[7] BS; Christine Schammel PhD; Kevin Hutson PhD; Justin Collins, BS; Jordan Brown and MD W. Jeffery Edenfield. Meaningful analysis of small data sets: A clinician's guide. *Greenville health system*, 2016.

[8] Anastasiya Motrenko, Vadim Strijov, and Gerhard-Wilhelm Weber. Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics*, 255:743 – 752, 2014.

[9] Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R. Golub, and Jill P. Mesirov. Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, 10(2):119–142, apr 2003.