Система эмпирического измерения качества алгоритмов классификации

Воронцов К. В., Инякин А. С., Лисица А. В.
voron@ccas.ru, inyakin@forecsys.ru, lisitsa@forecsys.ru
Москва, Вычислительный Центр РАН, ЗАО «Форексис»

Тестирование в режиме скользящего контроля является стандартной методикой сравнения алгоритмов классификации. Существующие системы поддержки научных исследований MatLab, R, DELVE, WEKA, и др. позволяют проводить такое тестирование, но требуют от пользователя достаточно высокой квалификации. На основе технологии, описанной в [2], авторами разрабатывается открытая распределённая система AxTTA (читается «акста», от Algorithms × Tasks Testing Area — полигон для тестирования алгоритмов на задачах), предоставляющая доступ к задачам, методам и результатам тестирования через интуитивно понятный web-интерфейс. Система предназначена как для специалистов по анализу данных, так и для экспертов-прикладников. Она позволяет полностью автоматизировать типовое исследование, цель которого — выяснить, какой из известных методов лучше подходит для решения конкретной прикладной задачи классификации или класса задач.

Набор задач взят из общедоступного репозитория UCI [3] и может пополняться пользователями. Алгоритмы запускаются на удалённых вычислительных серверах. Формирование выборок данных и оценивание качества классификации производится центральным сервером. Результаты тестирования представляются в виде таблицы «алгоритмы × задачи». Пользователь может формировать наборы отображаемых алгоритмов и задач, просматривать исходные данные задач, задавать управляющие параметры алгоритмов, назначать состав информации, отображаемой в ячейках таблицы. Обычно при сравнительном анализе алгоритмов классификации в таблицу выводятся только оценки скользящего контроля. Система AxTTA для каждой ячейки вычисляет расширенный набор критериев, позволяющий глубоко исследовать особенности как алгоритмов, так и задач. В сообщении рассматриваются некоторые методологические аспекты системы AxTTA.

Класс решаемых задач. Пусть X — множество объектов, Y — конечное множество имён классов, $X^L = \{(x_i, y_i)\}_{i=1}^L \subset X \times Y$ — выборка длины L. Объекты x из X описываются признаками $f_1(x), \ldots, f_n(x)$, возможно, разнотипными. *Задача классификации* задаётся $L \times n$-матрицей данных $\big[f_{ij}\big] = \big[f_j(x_i)\big]$ и целевым вектором $\big[y_i\big]$. Дополнительно может быть задана матрица потерь $\big[C_{yy'}\big]$, где $C_{yy'}$ — штраф за отнесение объекта класса y к классу y', а также некоторая априорная информа-

ция о признаках. Матрица данных может содержать пропуски. Требуется построить алгоритм классификации $a\colon X \to Y$, аппроксимирующий неизвестную целевую зависимость $y(x)$ на всём множестве X.

Метод обучения μ по обучающей выборке $X^\ell \subseteq X^L$ строит алгоритм классификации $a = \mu(X^\ell)$.

Качество алгоритма a на конечной выборке U характеризуется *частотой ошибок* $\nu(a,U) = \frac{1}{|U|}\sum_{x\in U}\big[a(x) \neq y(x)\big]$.

Процедура скользящего контроля является основой для вычисления большинства критериев. Производится N разбиений выборки X^L на обучающую подвыборку длины ℓ и контрольную длины k, $X^L = X_n^\ell \cup X_n^k$, $L = \ell + k$, $n = 1,\ldots,N$. *Оценка скользящего контроля* для функции $\xi\colon \{1,\ldots,N\} \to \mathbb{R}$ определяется как среднее $\hat{\mathsf{E}}\xi = \frac{1}{N}\sum_{n=1}^{N}\xi(n)$. Разбиения строятся по стандартной методике $t\times q$-fold cross-validation [5]: генерируется t случайных разбиений выборки X^L на q блоков примерно равной длины и равными долями классов, и каждый блок поочерёдно становится контрольной выборкой. Таким образом, $N = tq$ и $k = \frac{L}{q}$ с точностью до округления.

Качество классификации на n-м разбиении характеризуется частотой ошибок на обучении $\nu_n^\ell = \nu(a_n, X_n^\ell)$ и на контроле $\nu_n^k = \nu(a_n, X_n^k)$, где $a_n = \mu(X_n^\ell)$. Обобщающая способность метода μ на выборке X^L характеризуется одной из *оценок скользящего контроля*

$$\mathrm{CV}(\mu, X^L) = \hat{\mathsf{E}}\nu_n^k; \qquad \mathrm{CV}_\varepsilon(\mu, X^L) = \hat{\mathsf{E}}\big[\nu_n^k - \nu_n^\ell > \varepsilon\big].$$

Графики зависимости CV и CV_ε от ℓ при фиксированном k позволяют оценивать достаточную длину обучения для данного метода в данной задаче. График $\mathrm{CV}_\varepsilon(\varepsilon)$ позволяет оценивать риск переобучения.

Разложение ошибки на вариацию и смещение (bias-variance decomposition) [5]. Введём функцию *среднего предсказания*

$$\tilde{y}(x) = \arg\max_{c\in Y} \hat{\mathsf{E}}\big[a_n(x) = c\big], \quad x \in X.$$

Назовём $B(x) = \big[\tilde{y}(x) \neq y(x)\big]$ *смещением* метода μ на объекте $x \in X$. Соответственно, объекты выборки X^L разделятся на смещённые и несмещённые. Для произвольной конечной выборки $U \subset X$ определим среднее смещение $B(U) = \frac{1}{|U|}\sum_{x\in U} B(x)$. Тогда имеет место разложение:

$$\mathrm{CV}(\mu, X^L) = \hat{\mathsf{E}}B(X_n^k) + V(\mu, X^L),$$

где первое слагаемое характеризует смещённость модели классификации, используемой в методе μ; второе слагаемое, называемое *вариацией*, характеризует изменчивость результата обучения по отношению к составу

обучающей выборки. Если первое слагаемое велико, то надо менять саму модель. Если второе слагаемое велико, то качество классификации можно улучшить путём регуляризации или композиции алгоритмов.

Профиль устойчивости показывает, насколько изменяются классификации получаемого алгоритма, если состав обучающей выборки изменяется на m объектов:

$$S_m(\mu, X^L) = \hat{\mathsf{E}}_n \hat{\mathsf{E}}_{n'} \frac{1}{|X_{nn'}|} \sum_{x \in X_{nn'}} \big[\rho(X_n^\ell, X_{n'}^\ell) = m\big] \big[a_n(x) \neq a_{n'}(x)\big],$$

где $m = 1, \ldots, \min\{\ell, k\}$, $X_{nn'} = X_n^k \cap X_{n'}^k$, $\rho(U, V)$ — число несовпадающих объектов в выборках U и V. График профиля устойчивости, как правило, монотонно возрастает. Чем ниже проходит начальный (левый) участок профиля, тем устойчивее обучение.

Профиль разделимости определён для вещественнозначных алгоритмов классификации вида $a(x) = \arg\max_{c \in Y} \Gamma_c(x)$, где $\Gamma_c(x)$ — оценка принадлежности объекта x классу c. Степенью граничности или *отступом* (margin) объекта $x_i \in X^L$ называется величина [4]

$$M(a_n, x_i) = \Gamma_{n,y_i}(x_i) - \max_{c \in Y \setminus y_i} \Gamma_{n,c}(x_i).$$

Отступ показывает, насколько близко объект x_i подходит к границе класса y_i. Если объект оказывается за границей, то отступ отрицателен, и на данном объекте алгоритм допускает ошибку. Чем больше отступы, тем лучше качество классификации [4]. Эмпирические распределения отступов (*профили разделимости*) на обучающих и контрольных данных показывают, насколько надёжно данный метод разделяет классы.

Профиль представительности объектов. Для каждого объекта x_i из X^L вычисляется доля разбиений, при которых данный объект попадает в контроль, и на нём допускается ошибка:

$$I_i(\mu, X^L) = \hat{\mathsf{E}}\big[x_i \in X_n^k\big] \big[a_n(x_i) \neq y_i\big], \quad i = 1, \ldots, L.$$

Упорядоченная последовательность значений $I_1 \geqslant \ldots \geqslant I_L$ называется *профилем представительности* объектов. В начальный участок профиля попадают *шумовые* объекты, для которых $I_i > 0.5$. Объекты, оказавшиеся шумовыми для многих методов, по всей видимости, объективно являются таковыми. Система `AxTTA` позволяет отбросить шумовые объекты и провести обучение только по представительным объектам.

Профиль информативности признаков — это зависимость CV от числа использованных признаков. Отбор признаков осуществляется последовательным отбрасыванием наименее значимых признаков. Эта

процедура подходит для любого метода обучения, но ресурсоёмка, что, впрочем, вполне приемлемо в системе распределённых вычислений.

Временны́е показатели работы алгоритма. Поскольку в системе AxTTA обучение производится многократно на подвыборках разной длины, появляется возможность аппроксимировать время обучения зависимостью вида $T(\ell) = T_0 \ell^{\alpha} (\ln \ell)^{\beta}$, где коэффициенты α и β показывают эффективность алгоритма, а множитель T_0 зависит от параметров вычислительного сервера и не представляет интереса.

Сравнительный анализ методов для фиксированной задачи. Система AxTTA позволяет провести сравнительный анализ заданного пользователем набора методов по заданному набору критериев. При этом соответствующие разным методам таблицы сводятся вместе, а графики — накладываются и изображаются в одном масштабе.

Работа выполнена при поддержке РФФИ, проекты №№ 07-07-00372, 07-07-00181, 05-07-90410, 05-01-00877, а также программы ОМН РАН Алгебраические и комбинаторные методы математической кибернетики.

Литература

[1] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. — 2004. — № 13. — С. 5–36.

[2] *Качалков А. В., Хачай М. Ю.* Квазар-Оффлайн. Распределенный вычислительный комплекс для решения задач распознавания образов // ММРО-13 (в настоящем сборнике). — 2007. — С. ??–??.

[3] *Asuncion A., Newman D. J.* UCI Machine Learning Repository — University of California, Irvine. — 2007. — www.ics.uci.edu/~mlearn/MLRepository.html.

[4] *Garg A., Roth D.* Margin distribution and learning algorithms // Int. Conf. on Machine Learning (ICML'03), Washington, DC USA. — 2003. — Pp. 210–217.

[5] *Webb G. I.* MultiBoosting: A technique for combining boosting and wagging // Machine Learning. — 2000. — Vol. 40, No. 2. — Pp. 159–196.