Прикладные задачи в анализе данных

Домашнее задание 2

Попов Дмитрий Олегович,
517 группа

1 Задача и результат исследования

Задача: верно ли, что максимальное значение точности (т.е. значение точности при оптимальном выборе порога) всегда не меньше максимального значения сбалансированной точности?

Ответ: нет, неверно, приводятся примеры для всех отношений (больше, меньше и равно) и как автор пришёл к ним.

2 Условности и обозначения

Считаем, что объекты представлены действительными числами от 0 до 1 и каждый принадлежит ровно одному из классов нулевого и первого. Для упрощения доказательств вместо дискретных объектов использовались плотности классов, но результат можно сколь угодно хорошо приблизить конечными множествами. Поскольку неравенства будут получены строгие, существуют такие конечные множества, которые их выполняют.

Обозначения:
$f : [0, 1] \to \mathbb{R}_{\geq 0}$ — плотность объектов нулевого класса,
$g : [0, 1] \to \mathbb{R}_{\geq 0}$ — плотность объектов первого класса,
$f_0,\ f_1,\ g_0,\ g_1$ — мера множества объектов класса, соответствующего букве, отнесённых алгоритмом к классу, соответствующему индексу,
$ACC : [0, 1] \to [0, 1]$ — accuracy в зависимости от порога,
$BA : [0, 1] \to [0, 1]$ — balanced accuracy в зависимости от порога.

3 Примеры

3.1 Accuracy < Balanced accuracy

$f(x) = \sqrt{2}$,
$g(x) = 2 * \mathbb{1}[x > \frac{1}{2}]$.
Таким образом, нулевой класс представлен в $\sqrt{2}$ раз больше, чем первый, но распределён равномерно по всему единичному отрезку, а первый класс — равномерно по его второй половине.
Покажем, что оптимальный порог равен $\frac{1}{2}$:

$$ACC(\frac{1}{2}) = \frac{\sqrt{2} * \frac{1}{2} + 1}{\sqrt{2} + 1}$$

$$\tau < \frac{1}{2} \implies ACC(\tau) = \frac{\sqrt{2} * \tau + 1}{\sqrt{2} + 1} < ACC(\frac{1}{2})$$

$\tau > \frac{1}{2} \implies ACC(\tau) = \frac{\sqrt{2} * \tau + 2 * (1-\tau)}{\sqrt{2}+1} = ACC(\frac{1}{2}) + \frac{1 - 2\tau + \sqrt{2}(\tau - \frac{1}{2})}{\sqrt{2}+1}$,
$(\tau > \frac{1}{2} \iff \frac{1 - 2\tau + \sqrt{2}(\tau - \frac{1}{2})}{\sqrt{2}+1} l0) \implies ACC(\tau) < ACC(\frac{1}{2})$.
Мы вычислили оптимальную точность. Посчитаем значение сбалансированной точности при этом пороге:
$BA(\frac{1}{2}) = \frac{\frac{1}{2} + \frac{1}{1}}{2} = 0.75$.
$ACC(\frac{1}{2}) = \frac{\sqrt{2} * \frac{1}{2} + 1}{\sqrt{2}+1} \approxeq 0.707 \implies ACC(\frac{1}{2}) < 0.71 < 0.75 = BA(\frac{1}{2})$.

3.2 Accuracy = Balanced accuracy

Действительно легко заметить, что при сбалансированных классах сбалансированная точность вырождается в обычную, в этом случае равенство достигается при любых порогах и любых плотностях классов.

3.3 Accuracy > Balanced accuracy

Типичный случай при несбалансированных классах. Приведём достаточно простой пример:
$f(x) = 2 - 2x$,
$g(x) = x$.
Эти функции непрерывные, монотонные и пересекаются $\implies$ можно показать, что эта единственная точка пересечения будет являться оптимальным порогом для точности. Для этого нужно продифференцировать точность по порогу и заметить, что производная равна нулю при равенстве функций плотности.
$f(\tau) = g(\tau) \implies 2 - 2\tau = \tau \implies \tau = \frac{2}{3}$.
$ACC(\tau) = \frac{\frac{8}{9} + \frac{5}{18}}{1 + \frac{1}{2}} = \frac{7}{9}$,
$BA(\tau) = \frac{\frac{8}{9} * \frac{1}{1} + \frac{5}{18} * \frac{1}{2}}{2} = \frac{37}{72}$,
$ACC(\tau) - BA(\tau) = -\frac{19}{72} < 0$.

4 Подход к решению

Для начала поймём, как из порога и функций плотности выражаются обычная и сбалансированная точности:

$$ACC(\tau) = \frac{TN + TP}{N + P} = \frac{\int_0^\tau f(x)dx + \int_\tau^1 g(x)dx}{\int_0^1 (f(x) + g(x))dx},$$

$$BA(\tau) = \frac{\frac{TN}{N} + \frac{TP}{P}}{2} = \frac{\frac{\int_0^\tau f(x)dx}{\int_0^1 f(x)dx} + \frac{\int_\tau^1 g(x)dx}{\int_0^1 g(x)dx}}{2}.$$

Чтобы отклонить сбалансированную точность от обычной, нужно сделать размеры классов разными. Постараемся намеренно сделать сбалансированную точность выше, для этого положим:

$N = K * P, TP = K * TN, K > 0.$

В таком случае:

$$BA = \frac{\frac{TN}{K * P} + \frac{K * TN}{P}}{2},$$

$$ACC = \frac{TP + TN}{P + N} = \frac{(K + 1)TN}{(K + 1)P} = \frac{TN}{P}.$$

При $K \geq 2$ получим превышение сбалансированной точности над обычной.

Для простоты представим классы пороговыми функциями, то есть,

$f(x) = F_1 * \mathbb{1}[x < F_2],$

$g(x) = G_1 * \mathbb{1}[x > G_2],$

где F_1, F_2, G_1, G_2 — высота и смещения пороговых функций. Будем выбирать параметры так, чтобы плотность первого класса была выше плотности нулевого и они не были равны нулю одновременно, в таком случае окажется, что оптимальный порог находится в точке G_2.

Для завышения сбалансированной точности постараемся поместить как можно больше плотности нулевого класса правее порога и как можно меньше плотности первого класса — левее. В таком случае получаем $F_2 = 0, G_2 = \tau$. Для определённости положим $\tau = \frac{1}{2}, G_1 = 2$. В таком случае имеем $TP = P, K^2 = \frac{1}{\tau}$. После подстановки всех значений и проверки неравенства автор получил первый пример из этого файла.