Мультиклассовая классификация (Multiclass classification)

> Синонимы: Мультиноминальная классификация
> Разделы: Алгоритмы

В машинном обучении мультиклассовой называют задачу классификация, в которой метка класса принимает более чем два значения. В анализе данных наиболее часто реализуется именно этот вид классификации.

Задача формулируется следующим образом. Пусть задан обучающий набор данных $S = (x_i, y_i), i = 1, \ldots, n$, где n — число примеров, x_i — вектор признаков, y_i — метка класса i-го примера. Требуется построить классификатор, который для каждого x будет правильно предсказывать y, причем не только на обучающем множестве S, но и для любых наблюдений, в него не входящих (т.е. классификатор должен обладать обобщающей способностью).

В машинном обучении используется множество алгоритмов и методов классификации. Некоторые из них поддерживают возможность работы с несколькими классами естественным образом (нейронные сети, деревья решений, метод k-ближайших соседей и т.д.). Другие — бинарные, т.е. могут решать задачи классификации только для двух классов (логистическая и пробит регрессия, машины опорных векторов, линейный дискриминантный анализ и т.д.).

При этом задачу мультиклассовой классификации можно свести к нескольким задачам бинарной, что в некоторых случаях позволяет получить более точное и простое решение. Кроме этого, алгоритмы бинарной классификации более разработаны и математически обоснованы, в то время как мультиклассовые являются эвристиками.

Существуют несколько методов, которые позволяют преобразовать мультиклассовую задачу в набор бинарных.

Один против всех (One-versus-all, OvA или один против остальных, One-versus-rest, OvR). Для каждого класса строится один бинарный классификатор. При этом примеры класса определяются как «положительные», а всех других — как «отрицательные». Итоговый результат формируется по принципу «победитель получает все»: объект будет отнесен к классу, для которого бинарный классификатор даст большее число «положительных» примеров.

Метод имеет недостаток, что обычно каждый бинарный классификатор обучается в условиях дисбаланса классов, что снижает точность.

Один против одного (One versus One, OvO). Строится $k(k-1)$ классификаторов, позволяющих различить любую пару примеров разных классов. Алгоритм просматривает все пары примеров с разными метками классов и для каждой решает бинарную задачу f_{ij}. В каждом случае для пар (i, j) положительные — все примеры с метками i, а отрицательными — с j. Решение при этом имеет вид:

$$\hat{y} = \underset{x \in X}{argmax} \sum_{i} f_{ij}(x).$$

Недостатком метода является высокая трудоемкость: число классификаторов растет квадратично к числу примеров, в то время как у метода «один против всех» зависимость линейная.

Метод корректирующих кодов (Error-Correcting Output Codes — ECOC). Позволяет сократить число классификаторов с k (как в методе OvA) до $log_2 k$. Каждый класс кодируется в виде битовой последовательности, называемой кодовым словом. Однако при наличии даже одного некорректного бита в нем метка класса будет неверной. Чтобы избежать этого, вводится избыточность в виде нескольких дополнительных битов, называемых **корректирующими**.

Пусть каждый класс c_i $(i = 1, .., k)$ связан с кодовым словом $w_i \in \{0, 1\}^n$ длиной n. Обозначим j-й бит i-го кодового слова b_{ij}. Тогда набор кодовых слов можно представить в виде кодовой матрицы $m_{ij} \in \{0, 1\}^{k \times n}$, где каждая i-я строка описывает кодовое слово w_i, а столбец соответствует бинарному классификатору f_j. Множество классификаторов обозначим как $C = (f_1, f_2, \ldots, f_n)$.

Кодовая матрица, таким образом, описывает схему исходной мультиклассовой задачи. В каждом j-м столбце i-я строка содержат 1 для тех классов, обучающие примеры которых используются как положительные, и 0 — для тех, которые считаются отрицательными для данного классификатора f_j. Например, для задачи с 4-мя классами и 6-ю классификаторами, кодовая матрица может иметь вид:

Класс	f_1	f_2	f_3	f_4	f_5	f_6
1	1	1	1	0	0	0
2	1	0	0	1	1	0
3	0	0	0	1	0	1
4	0	0	1	0	1	1

Из таблицы видно, что первый классификатор использует классы 1 и 2 как положительные примеры, а для классов 3 и 4 — как отрицательные.

в процессе классификации используются все бинарные классификаторы, которые совместно формируют n-мерный вектор предсказаний. Он декодируется в одно из исходных значений классов, например, путем присвоения объекту того класса, кодовое слово которого наиболее близко к предсказанному вектору.

Таким образом, для примера x все бинарные классификаторы формируют предсказания, образующие вектор $y = (f_1(x), f_2(x), \ldots, f_n(x))$, который сравнивается с кодовыми словами для классов. Класс $\hat{c}_i$, кодовое слово которого окажется наиболее близким к y в смысле некоторой метрики $d(\cdot)$, и будет служить общим предсказанием мультиклассового классификатора:

$$\hat{c} = \underset{c}{\operatorname{argmin}} d(w_c, y).$$

Мерой близости между двоичными векторами может служить расстояние Хемминга, определяемое как число битовых позиций, в которых предсказанный вектор y отличается от кодового слова класса w_i, т.е.

$$d_h(w_i, y) = \sum_{j=1}^{n} |m_{ij} - y_j|.$$

Число классификаторов превосходит количество классов, т.е. $m > k$, что позволяет использовать более длинные кодовые слова. Поэтому сопоставление предсказанного вектора не будет искажено ошибками отдельных бинарных классификаторов.

Таким образом, метод корректирующих кодов не только позволяет сводить сложные мультиклассовые задачи классификации к набору бинарных, но и позволяет добиться более высокой точности.

Полиномиальная логистическая регрессия. Использует для преобразования бинарной классификации к мультиклассовой логистическую регрессию. Она является бинарным классификатором, формирующим на выходе рейтинг, изменяющийся в диапазоне от 0 до 1. Он может быть интерпретирован как вероятность принадлежности к «положительному» классу.

Для этого используется дискриминационный порог: если рейтинг выше его значения, то объект относится к «положительному» классу, в противном случе — к «отрицательному».

В основе работы модели лежит функция, называемая **softmax** обобщение логистической функции для многомерного случая. Она преобразует вектор z размерности k в вектор той же размерности, каждый элемент в интервале $[0, 1]$, сумма которых равна 1. Элементы нового вектора интерпретируются как вероятности принадлежности объекта к соответствующему классу.

Softmax-регрессия — это алгоритм машинного обучения с учителем, используемый в задачах многоклассовой классификации. В отличие от обычной логистической регрессии, в нем используется не сигмоидальная функция активации $s(z)$, а векторная
$\Psi : \mathbb{R}^K \to (0,1)^K$

$$\Psi(z_1, z_2, \ldots, z_K) = \begin{bmatrix} \psi_1(z_1, z_2, \ldots z_K) \\ \psi_2(z_1, z_2, \ldots, z_K) \\ \ldots. \\ \psi_K(z_1, z_2, \ldots, z_K) \end{bmatrix},$$

где $\psi_k : \mathbb{R}^K \to (0,1)^K$ скалярная функция вида:

$$\psi_k(z_1, z_2, \ldots z_K) = \frac{exp(z_k)}{\sum\limits_{i=1}^{K} exp(z_i)}.$$

Несложно увидеть, что благодаря нормирующим свойствам знаменателя $0 < \psi_k(z_1, z_2, \dots z_K) < 1$. Кроме того, $\sum_{i=1}^{K} \psi(z_1, z_2, \dots, z_k) = 1$. Эти два свойства позволяют интерпретировать данную величину как вероятность i-го класса.

Более подробно с описанными методами можно ознакомиться в статье «Мультиклассовая классификация в машинном обучении».