Классификация по многим меткам (Multi-label classification)

> Синонимы: Многометочная классификация, Классификация с несколькими выходами, MLC, Multi-output classification
> Разделы: Алгоритмы

В анализе данных и машинном обучении классификацией по многим меткам (multi-label classification, MLC) называют тип задач, в которых наблюдения могут относиться к более чем одному классу одновременно. При этом классы не являются взаимоисключающими. Обучающие наборы данных для таких задач содержат несколько полей меток класса, и все они могут быть задействованы в процессе обучения. Количество возможных меток при этом формально не ограничено.

В обычной задаче классификации с единственной меткой класса (single-label classification, SLC) каждое наблюдение может принадлежать только одному из них. Поэтому классы являются взаимоисключающими: если объект принадлежит к одному из них, то он не может одновременно относиться ни к какому другому (или другим). Очевидно, что обучающий набор данных для SLC может содержать несколько потенциальных полей меток, но при обучении задействуется только одно из них.

Причины, по которым получила развитие MLC, связаны с неоднозначностью задач анализа, в которых объекты относятся одновременно к нескольким категориям. Например, статья в Интернете может быть посвящена спорту и политике, науке и социальной тематике, медицине и бизнесу, и т.д. Классификация по многим классам позволяет сделать анализ данных более информативным и наглядным.

Изображение тоже может принадлежать сразу нескольким категориям: горы и море, лес и равнина, и т.д. Клиент также может принадлежать сразу к нескольким сегментам — быть одновременно лояльным и высококонверсионным. Это знание позволит лучше понимать клиентов и совершенствовать взаимоотношения с ними.

В основе алгоритмической реализации MLC лежит **метод бинарной релевантности** (BR). Он позволяет преобразовать задачу с N метками к независимому обучению N моделей бинарной классификации, которые сопоставляют 0 или 1 каждой метке. При работе с новыми данными такой комбинированный классификатор присваивает наблюдениям метки классов, для которых были предсказаны единицы.

Рассмотрим пример с тремя классами в таблице.

	Наблюдения	Классы

Наблюдения	Классы
1	A, B
2	A
3	A, C
4	C
5	B
6	A

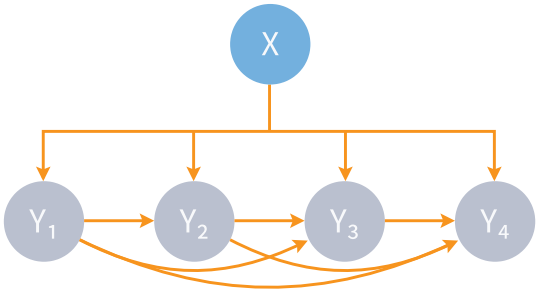
Для наблюдений 1 и 3 имеются две метки класса, для остальных — по одной. Применив метод бинарной релевантности, получим:

Классификатор	1	0
A	1, 2, 3, 6	4, 5
B	1, 5	2, 3, 4, 6
C	3, 4	1, 2, 5, 6

Если классификатор, связанный с классом A, выдаст 1, то наблюдения 1, 2, 3 и 6 принадлежат данному классу, а 4 и 5 — нет. Наблюдения 1 и 5, таким образом, относятся к классу B, а 3 и 4 — к C. Поскольку для 1-го наблюдения единицу дадут классификаторы A и B, следовательно, оно принадлежит обоим этим классам. Аналогично наблюдение 3 будет отнесено к классам A и C.

Очевидным недостатком метода является необходимость обучения числа классификаторов, равное количеству классов задачи, поэтому если их много, то это может привести к значительным вычислительным затратам. Преимуществом являются простота и интуитивная понятность.

Альтернативным подходом является использование **цепочек классификаторов** (classifier chain, CC). Здесь MLC также преобразуется к набору бинарных задач. Но в этом случае классификаторы работают последовательно, при этом выходы предыдущих используются в качестве входов для последующих.



Тогда первый классификатор в цепочке будет предсказывать класс A:

Наблюдение	Класс A
1	1
2	1
3	1
4	0
5	0
6	1

Следующий будет предсказывать класс B, но при этом результаты классификации для A также будут входными переменными.

Наблюдение	Класс A	Класс B
1	1	1
2	1	0
3	1	0
4	0	0
5	0	1
6	1	0

И, наконец, последний классификатор в цепочке будет предсказывать класс C.

Наблюдение	Класс A	Класс B	Класс C
1	1	1	0

Наблюдение	Класс A	Класс B	Класс C
2	1	0	0
3	1	0	1
4	0	0	1
5	0	1	0
6	1	0	0

При реализации метода следует учитывать, что для разных последовательностей классификаторов в цепи модель MLC будет давать разные результаты. Важным преимуществом метода CC по сравнению с BR является то, что он учитывает возможные зависимости между классами за счет использования результатов предсказания каждого класса для других.

Еще одним популярным методом MLC является **label powerset** (LP), который преобразует набор данных с несколькими метками в набор с несколькими классами, рассматривая каждую комбинацию меток как уникальный класс. Классификация по нескольким меткам достигается путем присвоения экземпляра классу, состоящему из набора меток.

Рассмотрим пример в таблице.

A	B	C	Класс
1	1	0	C110
0	0	1	C001
1	0	1	C101
0	1	1	C011

Затем классификатор обучается присваивать каждому наблюдению метку, указывающую на набор исходных классов, к которому он относится. Таким образом, метод LP преобразует задачу классификации со многими метками к обычной задаче с одной меткой.

На практике, поскольку MLC можно свести к задаче с одной меткой или к нескольким бинарным, любая аналитическая платформа, которая позволяет обучать соответствующие модели, может использоваться и для реализации сценариев классификации с несколькими метками.

Подробнее с решением задач классификации в аналитических платформах можно ознакомиться здесь: Классификация данных при помощи нейронных сетей, Классификация данных методом k-ближайших соседей.