Отбор признаков (Feature selection)

> Синонимы: Отбор переменных, Отбор атрибутов, Отбор факторов, Отбор предикторов, Генерализация, Variable selection, Attribut selection
> Разделы: Алгоритмы

Отбор признаков в статистике и машинном обучении — это процедура выбора из всего множества переменных, описывающих исследуемый бизнес-процесс, некоторого их подмножества, которое будет использоваться для построения аналитической модели. Является одним из типичных действий в процессе предобработки данных.

В основе идеи лежит предположение, что в анализируемых данных содержатся признаки, которые являются избыточными или нерелевантными (незначимыми). Избыточными являются те из них, которые коррелируют с другими и поэтому несут ту же информацию об исследуемом бизнес-процессе. Они могут быть исключены из рассмотрения без значительной потери информации, и, соответственно, точности модели.

Незначимыми считаются признаки, слабо или совсем не отражающие зависимости и закономерности в исходных данных, которые требуются обнаружить в процессе анализа.

Основными целями отбора являются:

- снижение размерности пространства признаков (избежание проклятия размерности);
- упрощение понимания и интерпретации модели, а также результатов анализа пользователем;
- сокращение времени обучения модели;
- согласование исходной информации с типом обучаемой модели (например, исключение номинальных атрибутов, если модель работает только с числовыми данными).

Кроме этого, правильный выбор подмножества признаков уже сам по себе может помочь выявить механизмы, лежащие в основе исследуемой проблемы, поэтому представляет самостоятельную ценность.

Практические задачи анализа могут содержать очень большое число признаков — десятки и даже сотни. При этом каждому их подмножеству соответствует своя функция, отражающая зависимость между ним и выходной переменной. Наиболее очевидной постановкой задачи отбора может быть построение моделей для всех подмножеств и выбора того из них, которое минимизирует частоту ошибок модели. Однако это решение реализуемо только для задач небольшой размерности.

Поэтому на практике наиболее часто используют три группы методов: методы-оболочки (оберточные), методы фильтрации и вложенные (встроенные).

Методы-оболочки предполагают, что для каждого подмножества признаков строится модель и оценивается на проверочном наборе данных. Этот метод является наиболее трудоемким, поскольку модель приходится строить для каждого набора, но при этом самым эффективным. В данном подходе выбранный набор является специфичным для модели определенного типа и оказывается неподходящим для другого.

Наиболее популярным методом этой группы является ступенчатый или пошаговый (stepwise) отбор. В его основе лежит последовательное добавление (forward selection) или исключение (backward selection) признаков, после каждого из которых проводится статистический тест для оценки значимости улучшения (при добавлении) или ухудшения (при исключении) точности модели.

Если при добавлении новой переменной в модель ее точность значимо возрастает, то ее включение целесообразно, в противном случае — нет. При последовательном исключении в модели сначала выбираются все признаки, а затем пошагово удаляются. Если при исключении очередного из них тест показывает значимое ухудшение модели (рост частоты ошибок), то его оставляют, в противном случае — исключают.

Недостатком подхода является трудоемкость: нужно строить много моделей и проводить статистических тестов. Кроме этого, поскольку проверка производится на основе одних и тех же данных, модели, построенные с помощью ступенчатого метода, склонны к переобучению.

При оценке значимости улучшения или уменьшения качества моделей применяются критерий Фишера (F-тест) или Стьюдента (t-тест), информационные критерии Акаике и байесовский.

Методы фильтрации вместо частоты ошибок используют некоторую быстро вычисляемую меру. Вычислительные затраты для методов фильтрации меньше, чем для оболочечных. При этом они не специфичны для типа модели. Недостатком подхода является снижение обобщающей способности.

Некоторые методы фильтрации предполагают ранжирование признаков с точкой отсечения, выбираемой с помощью перекрестной проверки.

Типичным примером метода фильтрации является использование коэффициента корреляции Пирсона, который показывает степень статистической связи между признаками и выходной переменной. Он вычисляется для каждого из них, затем задается некоторый дискриминационный порог (скажем, 0.8), и в модели оставляют только те, для которых значение коэффициента корреляции выше.

Вложенные методы являются частью алгоритма обучения самой модели, например, LASSO или Ridge-регрессия.

Отбор признаков может оказаться достаточно сложной как в алгоритмическом, так и вычислительном плане процедурой. Но тем не менее он является совершенно необходимым этапом моделирования, поскольку без правильно отобранного множества

значимых входных переменных построить хорошо работающую аналитическую модель невозможно.

Функционал, реализующий отбор признаков должен включаться в состав любой продвинутой аналитической платформы. Так, в Loginom отбор признаков используется при построении линейных регрессионных моделей: Отбор факторов в модели линейной регрессии.

Более подробно с технологией отбора признаков в аналитические модели можно ознакомится в статье Отбор переменных в моделях линейной регрессии.