



Нейросеть (классификация)

Описание

Компонент решает задачу классификации — в выходном наборе *Нейросеть* соотносит множество входных параметров (предикторов) с одним из заранее известных классов:

$\{ P(1), P(2), P(3) \dots P(n) \} \Rightarrow \text{Class}(i)$, где $P(n)$ — входной параметр, $\text{Class}(i)$ — один из известных классов.

Перед тем, как производить классификацию, алгоритм обучается на тренировочном наборе данных — обучающей выборке. Каждая строка такой выборки содержит:

- в полях, обозначаемых как *входные* — множество входных параметров;
- в единственном *выходном* поле — соответствующее этому набору обозначение класса.

Таким образом, перечень классов задается обучающим набором данных в процессе обучения *Нейросети* и не может быть изменен/пересмотрен в процессе классификации.

Технически обучение заключается в нахождении *весов* — коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными и выходными параметрами, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных и/или «зашумленных», частично искажённых данных. Для обучения используется квазиньютоновский метод Бройдена-Флетчера-Гольдфарба-Шанно с ограниченным использованием памяти *L-BFGS*.

Примечание: для каждого непрерывного параметра в структуре *Нейросети* будет создан один вход, в то время как для каждого дискретного – столько входов, сколько у данного параметра имеется различных уникальных значений.

Порты

Вход

-  Входной источник данных (таблица данных).

Требования к принимаемым данным

Поля входного набора, которые будут использоваться в качестве *входных* или *выходных*, не должны содержать пропущенные значения. Если это требование не выполнено, то в момент активации узла будет выдана ошибка.

Выход

-  Выход нейросети (таблица данных).
-  Сводка (переменные) — показатели качества модели.

Мастер настройки

Шаг 1. Настройка входных столбцов

На первом этапе необходимо задать назначение полей входного набора данных:

-  **Входное** — поле содержит значения одного из входных параметров.
-  **Выходное** — поле содержит значения классов.
-  **Не задано** — поле не участвует в обработке. Устанавливается по умолчанию для прочих полей.

Важно: назначение *Выходное* можно задать только для поля с *Дискретным видом данных*, вид данных *Входных* полей не регламентируется.

Шаг 2. Настройка нормализации

На этом этапе входные данные приводятся к нормальному виду — преобразуются из натуральных значений в безразмерные для того, чтобы данные, имеющие большой разброс значений, не преобладали над данными с меньшим разбросом значений. Использование нормализации увеличивает качество и скорость обучения *Нейросети*.

Шаг 3. Разбиение на множества

Страница *Разбиение на множества* позволяет разделить множество на обучающее и тестовое:

- Обучающее — структурированный набор данных, применяемый для обучения аналитических моделей. Каждая запись обучающего множества представляет собой обучающий пример, содержащий заданное входное воздействие и соответствующий ему правильный выходной (целевой) результат.

- Тестовое — подмножество обучающей выборки, содержащее тестовые примеры, т.е. примеры, использующиеся не для обучения модели, а для проверки его результатов.

Доступные параметры:

- Размер обучающего и тестового множества в процентах или строках.
- Метод разбиения на обучающее и тестовое множество. Существует три метода разбиения:
 - Случайный — случайно разбивает множество записей на обучающее и тестовое множество.
 - Последовательный — группы строк множеств (обучающее, неиспользуемое, тестовое) выбираются последовательно, т.е. сначала выбираются те записи, которые входят в первое множество, затем — во второе и т.д. Порядок множеств можно менять (кнопки *Сдвинуть вверх*, *Сдвинуть вниз*).
 - По столбцу — разбиение на обучающее и тестовое множества задаётся при помощи параметра. Параметром выступает столбец с логическим типом данных, где значение «ИСТИНА» указывает на то, что запись относится к тестовому набору, а значение «ЛОЖЬ» — на то, что запись принадлежит обучающему набору (т.е. можно разбить множество на обучающее и тестовое в узле Разбиение на множества и подать данные из порта *Общий выходной набор* на вход узла *Нейросеть (классификация)*, выбрав в качестве параметра разбиения по столбцу колонку "Тестовое множество"). При выборе данного метода таблица выбора соотношения обучающего и тестового множеств становится неактивной.
- Метод валидации, который может принимать следующие значения:
 - Без валидации.
 - K-fold кросс-валидация — позволяет выбрать *Метод сэмплинга* и количество *Колод кросс-валидации*.
 - Монте-Карло — позволяет выбрать *Количество итераций ресемплинга* и задать размер обучающего и валидационного множества.

Random seed — начальное число (целое, положительное), которое используется для инициализации генератора псевдослучайных чисел. Последовательность чисел генератора полностью определяется начальным числом. Если генератор повторно инициализируется с тем же начальным числом, он выдаст ту же последовательность чисел.

Параметр влияет на порядок случайного разбиения на тестовое и обучающее множество и на воспроизводимость результата обучения. Можно повторить результат обучения узла, если подать те же данные и выставить тот же random seed.

Для параметра доступны следующие команды:

- Всегда случайно — начальное число всегда будет случайным.
- Генерировать — сгенерируется новое начальное число.
- Копировать — в буфер обмена будет скопировано указанное значение.

Шаг 4. Настройка параметров Нейросети

Структура Нейросети

- **Количество скрытых слоев** — предоставляется выбор из списка:
 - Без скрытых слоев.

- Один скрытый слой (используется по умолчанию).
- Два скрытых слоя.
- **Количество нейронов в первом скрытом слое** — целое число ≥ 1 (по умолчанию = 10).
- **Количество нейронов во втором скрытом слое** — целое число ≥ 1 (по умолчанию = 10).

Параметры обучения

- **Количество рестартов** — число попыток обучения *Нейросети* (на одном и том же наборе), выполняемых из случайных начальных значений весов. По завершении всех рестартов выбирается сеть, которая обеспечивает наименьшую среднеквадратическую ошибку на обучающем множестве. Целое число ≥ 1 (по умолчанию = 10).
- **Коэффициент регуляризации** — коэффициент зависимости весов сети друг от друга. Чем больше эта зависимость, тем сильнее будет влияние одного входного параметра на другие. Регуляризация позволяет снизить эффективное число степеней свободы модели, избежав тем самым переобучения. Предоставляется выбор из следующих вариантов:

- Отсутствует:
 - 0
- Слабая:
 - 0,001. Используется по умолчанию.
 - 0,003
 - 0,010
- Средняя:
 - 0,030
 - 0,100
 - 0,300
- Сильная:
 - 1,000
 - 3,000
 - 10,000
 - 30,000
- Очень сильная:
 - 100,000

Обычно значение параметра выбирается из выпадающего списка. При необходимости можно самостоятельно задать значение в диапазоне от 0 до 100 .

- **Продолжить обучение** — установка данного флага позволяет начать переобучение модели не со случайных значений весов *Нейросети*, а с полученных при последнем обучении. При этом параметр *Количество рестартов* игнорируется.

Критерии останова

Обучение сети происходит итерационно. При каждой итерации считывается весь обучающий набор данных и изменяются веса *Нейросети*. Это продолжается до тех пор, пока относительные изменения весов не станут меньше заданного порога или количество итераций не превысит заданной величины.

- **Порог минимального изменения весов** — если на очередном шаге обучения относительное изменение нормы вектора весов становится меньше порога, то обучение останавливается. По умолчанию = 0,005.
- **Максимальное количество эпох** — максимальное количество итераций обучения алгоритма. Если процесс обучения необходимо ограничить по времени, в этом случае он остановится после заданного количества эпох, даже если обучение еще не пришло к оптимальной точке, т.е. не достигнут порог минимального изменения весов.

Шаг 5. Настройка автоматического подбора параметров Нейросети

Нейросеть имеет три подбираемых параметра, относящихся к структуре:

- Количество скрытых слоев (0, 1 или 2);
- Количество нейронов в каждом из скрытых слоев;
- **Коэффициент регуляризации** — параметр, регулирующий жесткость модели.

Общие параметры

- **Подобрать структуру** — автоматический подбор структуры *Нейросети*:
 - **Начать с указанной структуры** — использование в качестве начальных параметров значений, заданных на странице настройки параметров *Нейросети*.
- **Подобрать коэффициент регуляризации** — автоматический подбор коэффициента регуляризации *Нейросети*:
 - **Начать с указанного коэффициента регуляризации** — использование в качестве начального *Коэффициента регуляризации* значения, заданного на странице настройки параметров *Нейросети*.

Примечание: если необходимо осуществлять подбор параметров для больших входных объемов или сложных моделей, можно включить только подбор структуры, либо только подбор коэффициента регуляризации, сократив время на обучение.

Параметры выборки

Для ускорения процесса автоподбора предусмотрено задание подвыборки, на которой он будет производиться:

- **Использовать подмножество обучающего набора** — использование подвыборки обучающего множества для автоподбора.
 - **Размер выборки в процентах** — размер подвыборки обучающего множества.
 - **Максимальный размер выборки** — максимальный размер подвыборки обучающего множества.

Критерии останова автоподбора

По умолчанию процесс автоматического подбора останавливается при невозможности найти лучшие параметры, чем уже найденные. Для ограничения времени работы предусмотрена возможность ограничить, в том числе одновременно количество шагов автоподбора и время автоподбора:

- **Шагов автоподбора не более** — максимальное количество шагов алгоритма (0 — отключение ограничения);
- **Время автоподбора не более (сек.)** — максимальное время работы алгоритма (0 — отключение ограничения).

Примечание: при работе следует учитывать, что фактически оба ограничения могут быть незначительно превышены при использовании подвыборки для автоподбора, так как последним этапом, который не учитывается ограничениями, будет осуществлено обучение лучшей *Нейросети* на полном наборе.

Отдельный случай останова оптимизатора — если достигнут теоретически наилучший результат. Как для регрессионных сетей, так и для классификатора это равная 0 среднеквадратическая ошибка на обучающем множестве.

Для классификатора также по умолчанию включен *Останов при нулевой ошибке классификации*:

- **Останов при нулевой ошибке классификации** — прекращение автоподбора при достижении нулевой ошибки классификации.

Примечание: опцию *Останов при нулевой ошибке классификации* можно отключить, т.к. правильная классификация всех примеров не всегда означает наилучшую структуру *Нейросети*: оптимизатору можно дать возможность подобрать сеть с лучшей обобщающей способностью (например, с меньшим числом нейронов или сильнее регуляризованную), но при этом не обязательно с нулевой ошибкой классификации.

Стратегия оптимизации

Целевой функцией для оптимизатора является среднеквадратическая ошибка на обучающем наборе. При этом для учета случаев, когда несколько сетей показывают сравнимые по точности результаты, в целях выбора сети с самой простой структурой значение целевой функции дополнительно штрафует на слабо отличающийся от единицы множитель ($1+1e-8$) за каждый скрытый нейрон.

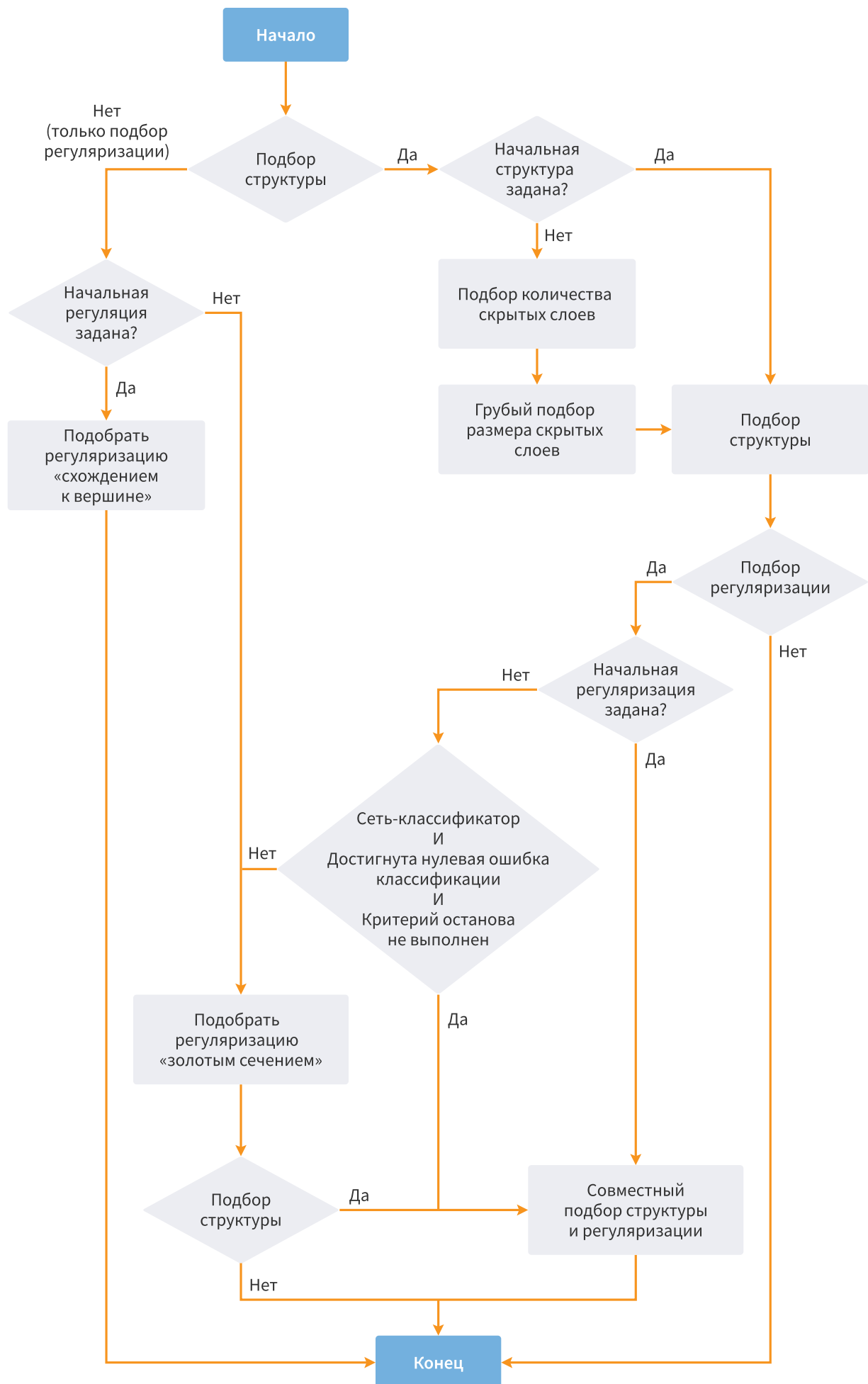
Стратегия оптимизации следующая:

- Если необходимо подобрать только коэффициент регуляризации для заданной структуры:
 - Если начальная точка не задана, то коэффициент регуляризации подбирается методом *золотого сечения*, в противном случае — методом *схождения к вершине*.
- Если необходимо подобрать только структуру, не изменяя коэффициент регуляризации:
 - Если не задана начальная структура, то она подбирается в два этапа: сначала происходит выбор количества скрытых слоев (0, 1 или 2), затем, если результат предыдущего этапа не 0,

грубо подбирается размер скрытых слоев методом *золотого сечения*, причем для 2 скрытых слоев количество нейронов на данном этапе делается одинаковым.

- Структура подбирается сразу по всем трем параметрам (число слоев, число нейронов) методом *схождения к вершине* из заданной либо подобранной начальной точки.
- Если необходим автоподбор структуры и регуляризации:
 - Подбирается структура так же, как и в предыдущем пункте. При этом, если начальное значение регуляризации задано, то используется оно, в противном случае — регуляризация отключена.
 - Если начальное значение регуляризации не было задано, оно подбирается методом *золотого сечения*.
 - Финальный этап автоподбора производится методом *схождения к вершине* по всем четырем параметрам.

Блок-схема (граф переходов) реализованной стратегии автоподбора указана на рисунке:



Статьи в разделе:

- [Выход нейросети](#)
- [Сводка](#)