# **ОПИСАНИЕ МОДЕЛИ**

## **БАЗОВЫЕ КОНЦЕПЦИИ**

В этом разделе представлены основные концепции и понятия, используемые в подходе динамического метода выбора регрессии. Эти концепции обеспечивают основу необходимую для понимания того, как работает данный метод.

В этой статье используются следующие математические обозначения:

* R = {R1, R2, …, Rn} – множество N базовых моделей регрессии,
* С = {C1, C2, …, Ck} - множество K базовых моделей классификации,
* X = {x1, x2, …, xn} – набор значений независимых атрибутов,
* y – целевая переменная,
* Xtrain – выборка значений независимых атрибутов, предназначенных для обучения моделей,
* ytrain – значения целевой переменной для всех наборов данных из выборки Xtrain,
* Xtest – выборка значени независимых атрибутов, предназначенных для проверки моделей,
* ypred – предсказанные значения целевой переменной для каждого набора значений атрибутов из выборки Xtest,
* ytest – фактические значения целевой переменной для всех наборов данных из выборки Xtest,
* MSE – среднеквадратичное отклонение
* SEi – квадрат отклонения i-того набора данных.

## **ОБЩЕЕ ОПИСАНИЕ МОДЕЛИ**

В первой главе были описаны различия между статическими и динамическими методами машинного обучения, и как уже было написано для задач классификации существуют множество алгоритмов динамического метода выбора классификаторов, но методов динамического выбора моделей регрессии не существуют. В данной работе описывается метод выбора модели регрессии на основе мета-обучения.

Как уже говорилось, все модели, использующие данные реального мира неидеальны и всегда имеют ошибки. Каждая регрессионная модель, применяемая на одном и том же наборе данных, дает разные ошибки. Конечно, могут быть случаи, когда предсказанные значения разных моделей регрессий будут одинаковыми, но в основном они отличаются. То есть для каждого объекта из набора данных можно выбрать ту модель, которая предсказала значение целевой переменной точнее всех. Исходя из этого, можно предположить, что для любого объекта из набора данных можно выбрать одну модель регрессии, которая предположительно будет давать лучший результат.

Цель динамического метода выбора модели регрессии состоит в том, чтобы исходя из обучающего набора предсказывать для каждого нового объекта набора данных метку модели регрессии, которая предположительно будет давать лучший результат, и предсказывать целевую переменную для данного объекта набора данных, использую выбранную модель регрессии.

## **ЭТАПЫ ПОСТРОЕНИЯ МОДЕЛИ**

Как и все модели машинного обучения, модель динамического метода выбора регрессии состоит из 2 основных этапов:

* этап обучения,
* этап предсказания,

после которых идет улучшение модели, с помощью манипуляций выбранных методов регрессий R = {R1, R2, …, Rn} и классификаций С = {C1, C2, …, Ck}.

### **ЭТАП ОБУЧЕНИЯ**

Первый этап построения модели — это обучение. Используя обучающие выборки необходимо сначала обучить все регрессионные модели, после чего предсказать значения целевой переменной для каждого объекта набора данных. Для того, чтобы избежать переобучения модели используется метод перекрёстного обучения, идея которого состоит в следующем: данные делятся на несколько равных частей, в нашем случае 5 частей, после чего каждый раз 4 из этих 5-ти частей используются для обучения моделей, а один для предсказания. Это повторяется 5 раз, и в итоге мы получаем предсказанные значения целевых переменных для всего обучающего набора данных.

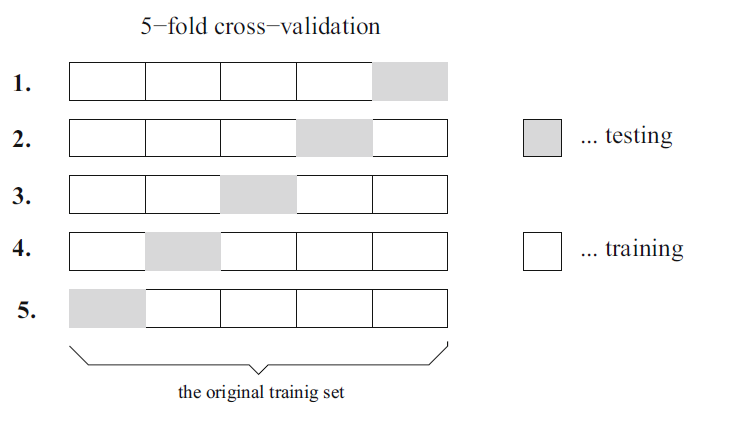


Рисунок 16. Способ перекрестного обучения моделей во избежание переобучения [Miroslav Kubat, 2015]

После того, как были получены предсказания целевых значений для всех регрессионных моделей, для каждой модели нужно посчитать квадрат отклонения предсказанного значения целевой переменной от реального значения для всех объектов используя следующую формулу:

На этом этапе определяются какие модели регрессии какие ошибки в предсказаниях выдали, что позволяет нам выбрать для каждого объекта данных ту модель, предсказанное значение которой дает наименьшую ошибку. Каждой модели регрессии присваивается метка. Создается новый атрибут, который хранит в себе для каждого объекта значение метки той модели регрессии предсказание которого дало наименьшую квадратную ошибку.

После следует этап обучения модели классификации, которая на вход получает все данные независимых переменных и атрибут с метками лучших регрессионных моделей как целевую переменную. Перед тем как закончить этап обучения все регрессионные модели обучаются еще раз, уже с полным набором данных на входе.

На этом этап обучения модели динамического метода выбора регрессионных моделей заканчивается, после которого следует этап предсказания.

|  |  |
| --- | --- |
| Алгоритм этапа обучения |  |
| def fit(R,Ci,(X,y)):  Вход: набор моделей регрессий R={R1, R2, ...,Rn}  модель классификации Ci  набор обучающих данных (X, y)  Выход: обученные модели регрессий и модель классификации  #для каждой модели из набора моделей регрессий  for model in R:  #обучение моделей с помощью перекрестного обучения  model.fit(X, y)  #предсказание целевых значений для набора X  y\_pred <- model.predict(X)  #расчет квадратичных ошибок для каждого объекта набора данных  err[model] <- (y - y\_pred)^2  #поиск лучшей модели для каждого атрибута с наименьшей ошибкой  for i in X.lenght:  #определение атрибута, который хранит индексы лучших моделей для  #каждого объекта  index <- model.index with idmin(err)  #обучение модели классификации  Сi.fit(X, index) | |

### **ЭТАП ПРЕДСКАЗАНИЙ**

Как написано выше, целью динамического метода выбора модели регрессии является выбор для каждого объекта данных лучшей модели регрессии из предложенного множества. Достичь этой цели нам помогает модель классификации, которая на этапе обучения обучается, используя полученные данные о лучших моделях регрессий для каждого объекта, после чего на этапе предсказания для каждого объекта нового набора данных предсказывает, какую из регрессионных моделей лучше использовать для предсказания. После того как модель классификации предсказывает метку регрессионной модели, выбранная регрессионная модель делает предсказание.

Таким образом мы для каждого объекта набора данных получаем свои предсказания исходя из выбранной модели регрессии классификатором.

|  |  |
| --- | --- |
| Алгоритм этапа предсказаний |  |
| def predict(X):  Вход: набор данных для предсказания X  Выход: предсказанные значения целевой переменной для набора X  #для каждого объекта из набора данных  for x in X:  #модель классификации предсказывает лучшую модель регрессии для  #объекта x  label <- Ci.predict(x)  #предсказание целевой переменной для объекта x с помощью  #найденной модели регрессии  yi <- R[label].predict(x)  #сохранение предсказанного значение yi в массиве  y\_pred.append(yi)  #возвращение значений целевых переменных для набора X  return y\_pred | |

### **ЭТАП УЛУЧШЕНИЯ МОДЕЛИ**

Так как изначально берется произвольное количество моделей регрессии и произвольная модель классификации, нужно понять правильно ли были выбраны модели регрессии и классификатор. Для этого нужно оценивать регрессионные модели используя перекрестную проверку (cross-validation) и как характеристики для сравнивания брать среднеквадратичную ошибку (RMSE) и стандартное отклонение.

В ходе обучения основной модели последовательно обучаются все используемые модели регрессий, и для каждого из них на выходе получаются значения RMSE и стандартного отклонения. Из всего набора регрессионных моделей выбираются те, которые по сравнению с остальными имеют лучшие результаты. Все остальные модели удаляются. Таким образом, удаляя плохие модели регрессий, предсказания основной модели улучшаются и значение RMSE снижается.

Следующим ходом улучшения модели, является выбор модели классификации. Модели классификаций сравниваются между собой с помощью значения показателя точности предсказаний. Полученная после удаления худших регрессионных моделей обучается, получая в ходе все модели классификаций пошагово, для каждого из которого определяется точность предсказания лучшей модели регрессии. Полученные результаты дают возможность выбрать лучший классификатор.

После получения лучших моделей регрессий и лучшей модели классификации модель готова к работе.