Вероятностный подход для задачи предсказания биологической активности ядерных рецепторов

Володин Сергей Евгеньевич

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 374, осень 2016

Цель исследования

Предсказание взаимодействия двух типов молекул: лиганд и рецепторов. Необходимо оценить вероятность связывания.

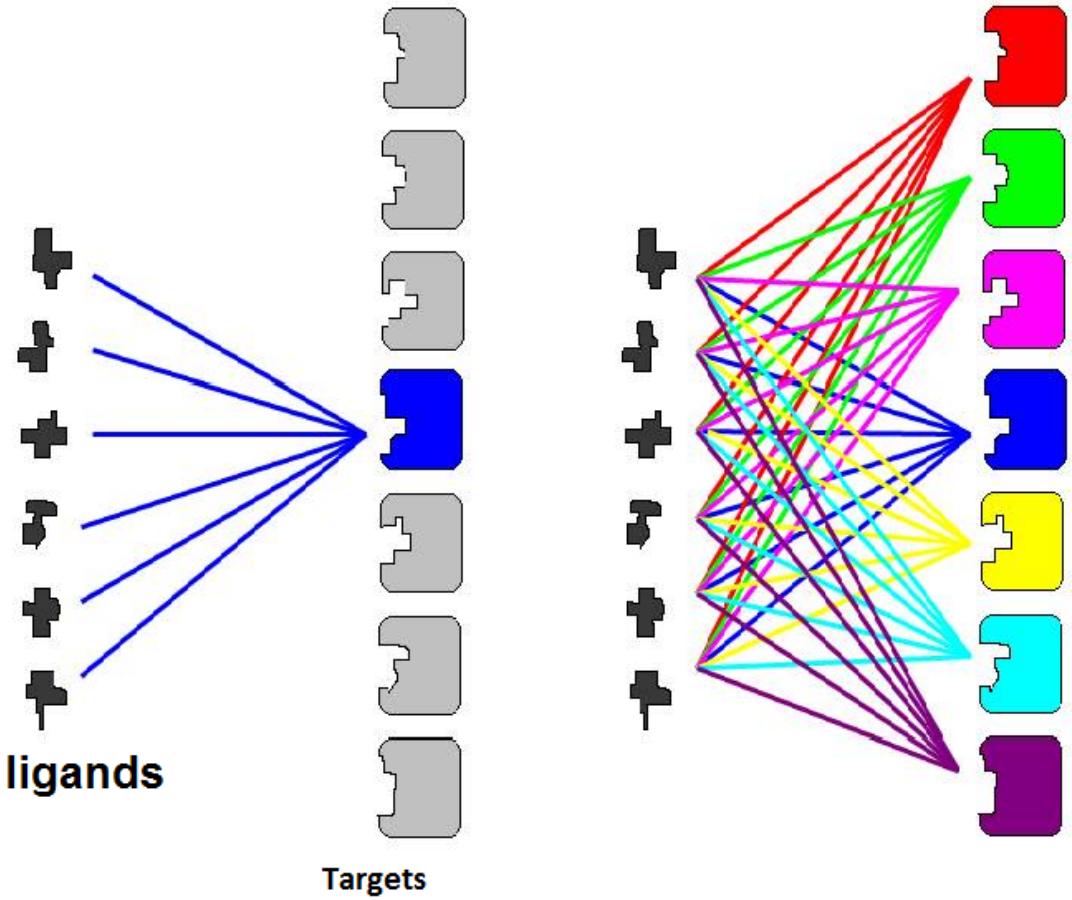
Проблема

События реакции лиганда с различными рецепторами не независимы.

Задача

Необходимо построить вероятностную модель, учитывающую зависимости между классами, а также построить бинарный классификатор.

Проблема



Литература

1. Olexandr Isayev Sherif Farag Stephen J. Capuzzi, Regina Politi and Alexander Tropsha. Qsar modeling of tox21 challenge stress response and nuclear receptor signaling toxicity assays.
2. Geoff Holmes Eibe Frank Jesse Read, Bernhard Pfahringer. Classifier chains for multi-label classification.
3. Eyke H.0 Krzysztof Dembczynski, Weiwei Cheng. Bayes optimal multilabel classification via probabilistic classifier chains. 2010.

Постановка задачи

Задана выборка $\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\} = \mathfrak{L} \sqcup \mathfrak{T}$. $\mathbf{x}_i \in \mathbb{R}^n$. $\mathbf{y}_i \in \{0, 1, \square\}^l$,
$\square$ — пропуск в данных.
$\mathbf{X}, \mathbf{Y}$ — случайные величины, между классами есть зависимости.
Модель классификации: функция $\mathrm{f}\colon \mathbf{W} \times \mathbf{X} \times \mathbf{Y} \to [0, 1]$,

$$\mathrm{f}(\mathbf{w}, \mathbf{x}, \mathbf{y}) = P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}; \mathbf{w})$$

Функция потерь — логарифм правдоподобия

$$Q(\mathrm{f}|\mathbf{w}, \mathcal{Z}) = -\sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{Z}} \log \mathrm{f}(\mathbf{w}, \mathbf{x}, \mathbf{y}) P(\mathbf{X} = \mathbf{x})$$

Требуется минимизировать Q:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbf{W}} Q(\mathrm{f}|\mathbf{w}, \mathfrak{L})$$

Для оценки конкретной модели используется AUC.

Решение

Вычислительный эксперимент

Цель эксперимента

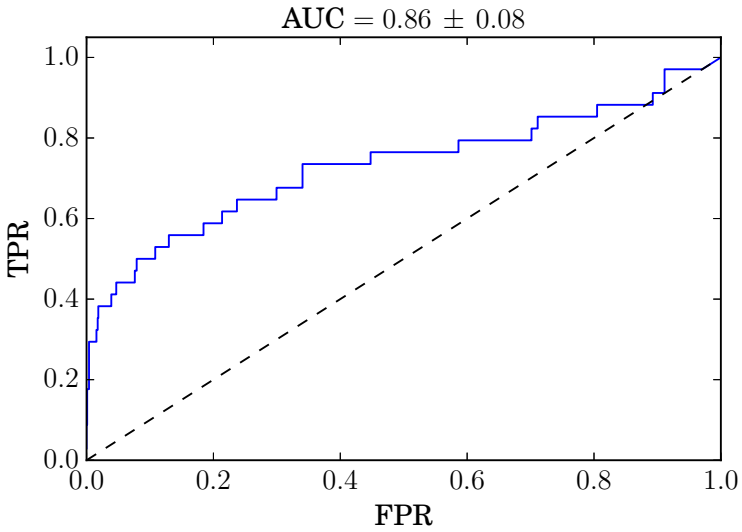
1. Сравнение различных моделей по критерию AUC для различных классов.
2. Выбор гиперпараметров исходя из внешних требований к решению задачи.

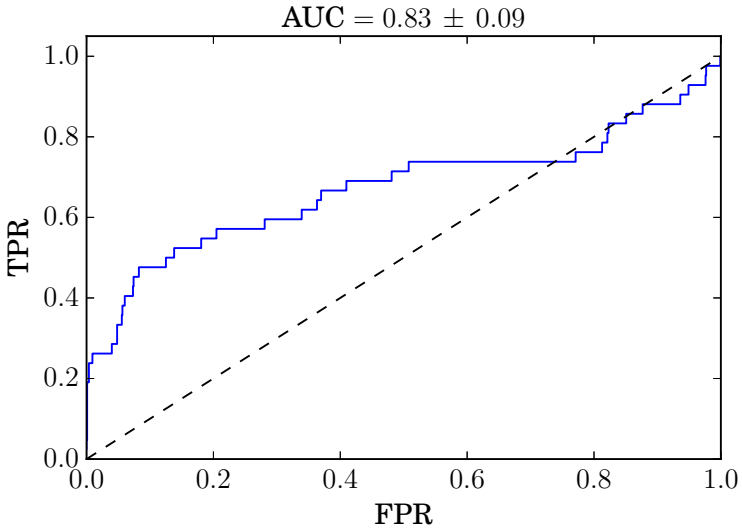
Сравниваемые модели

1. Binary Relevance
2. **PCC** — предлагаемое решение
3. Random Forest

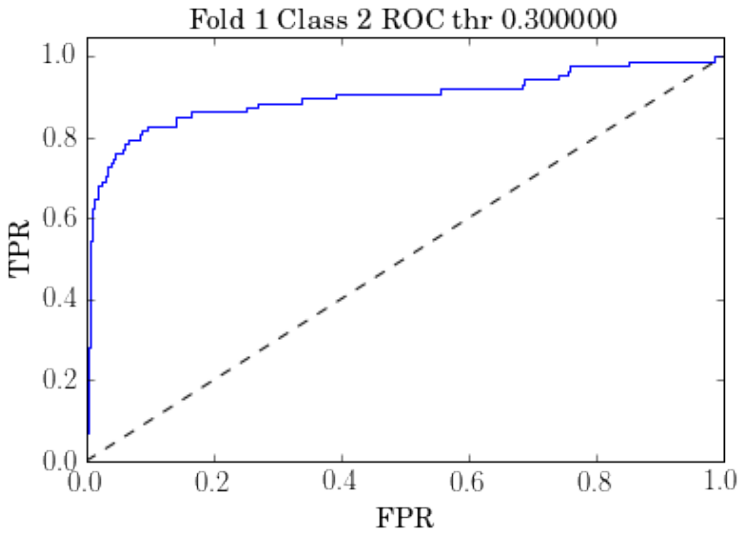
Вычислительный эксперимент

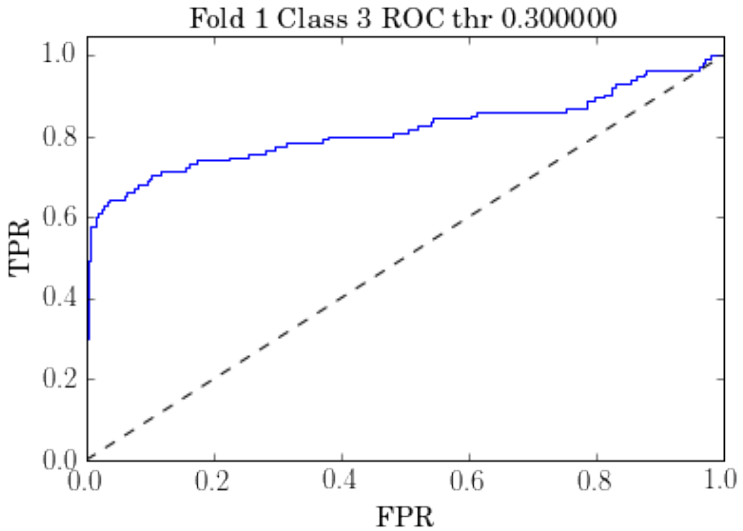
BR





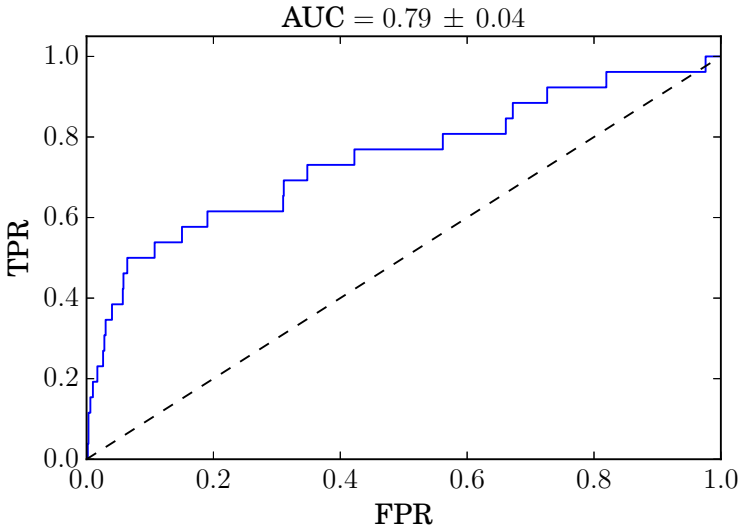
PCC

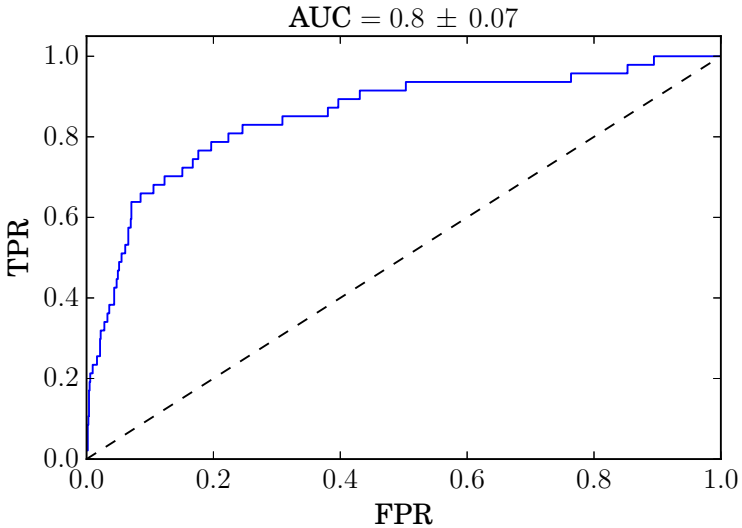




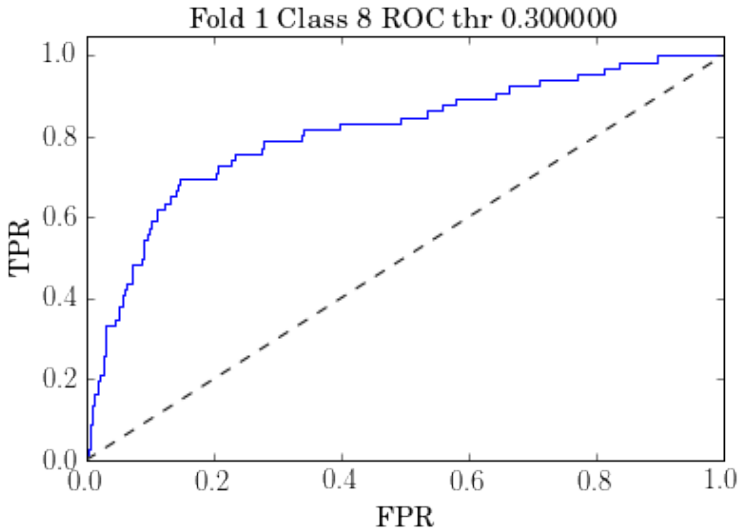
Вычислительный эксперимент

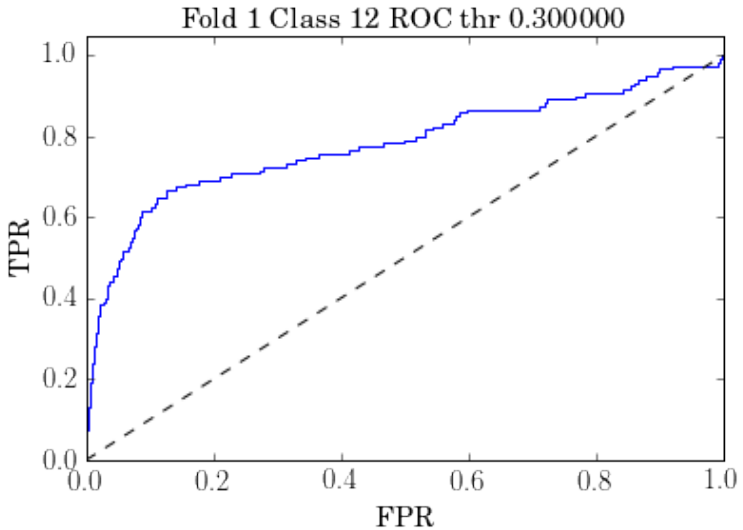
BR





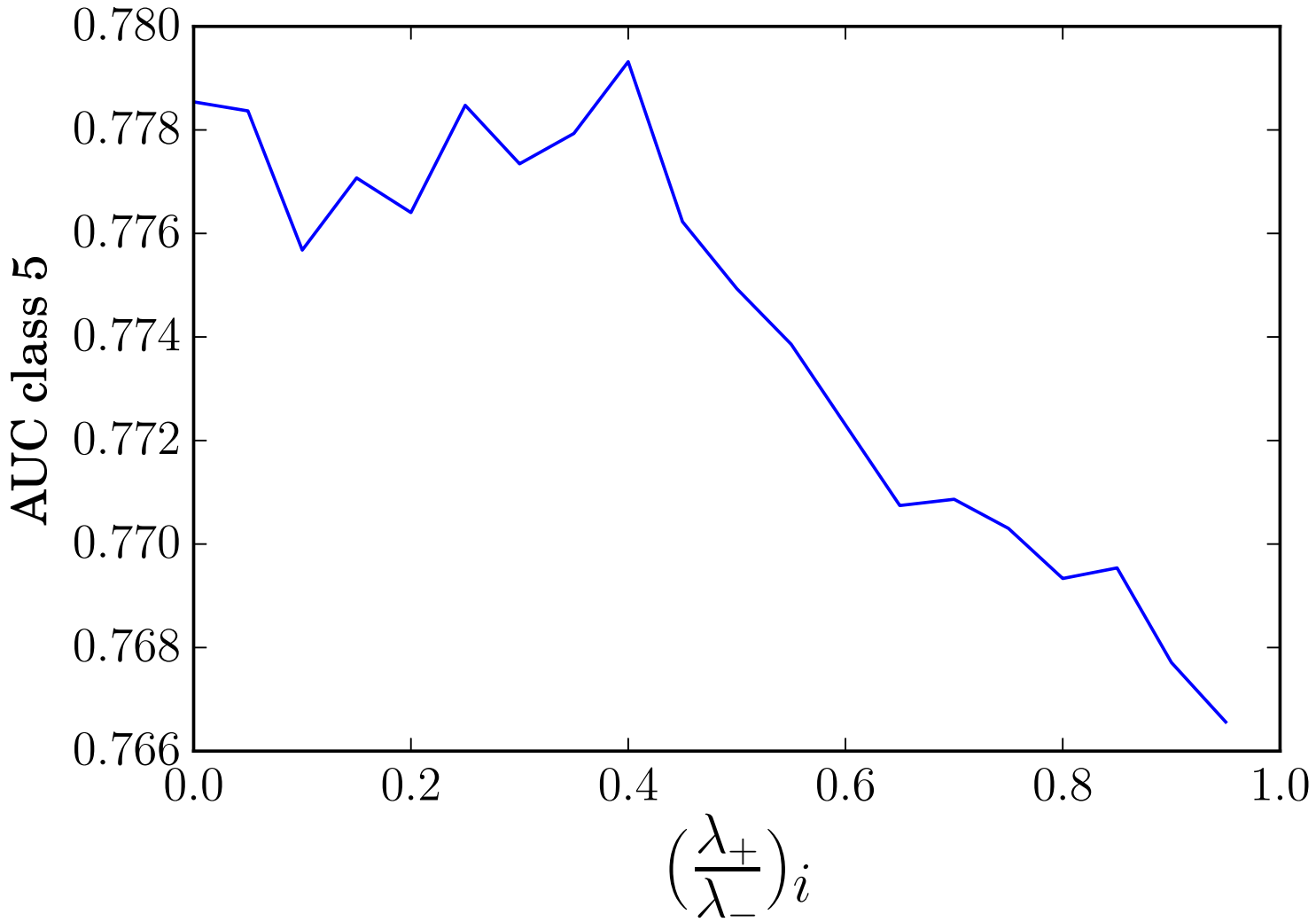
PCC





Вычислительный эксперимент

Зависимость от гиперпараметров



Рецептор	Binary Relevance	**PCC**	Random Forest
NR-AhR	**0.83** $\pm$ 0.03	**0.83**	**0.93**
NR-AR-LBD	**0.86** $\pm$ 0.08	**0.90**	**0.88**
NR-AR	**0.83** $\pm$ 0.09	**0.84**	**0.83**
SR-MMP	**0.87** $\pm$ 0.03	**0.87**	**0.95**
NR-ER	**0.78** $\pm$ 0.04	**0.78**	**0.81**
SR-HSE	**0.79** $\pm$ 0.04	**0.78**	**0.86**
SR-p53	**0.79** $\pm$ 0.07	**0.80**	**0.88**
NR-PPAR-gamma	**0.79** $\pm$ 0.04	**0.81**	**0.86**
SR-ARE	**0.78** $\pm$ 0.02	**0.78**	**0.84**
NR-Aromatase	**0.81** $\pm$ 0.05	**0.82**	**0.84**
SR-ATAD5	**0.81** $\pm$ 0.06	**0.80**	**0.83**
NR-ER-LBD	**0.80** $\pm$ 0.07	**0.82**	**0.83**

Заключение

1. Предложена модель для предсказания взаимодействия, учитывающая зависимости между классами
2. Проведено сравнение модели с другими по критерию AUC
3. Базовая модель BR имеет худшие показатели AUC, чем Random Forest
4. PCC лучше BR для большинства классов по критерию AUC при верных значениях гиперпараметров