

Вероятностный подход для задачи предсказания биологической активности ядерных рецепторов

Володин С. Е., Попова М., Стрижов В. В.
sergei.volodin@phystech.edu, maria_popova@phystech.edu, strijov@ccas.ru

Цель исследования

Предсказание взаимодействия двух типов молекул: лиганд и рецепторов. Задача является важной для разработки различных лекарств. Из-за нехватки данных биохимическое моделирование [1] неприменимо.

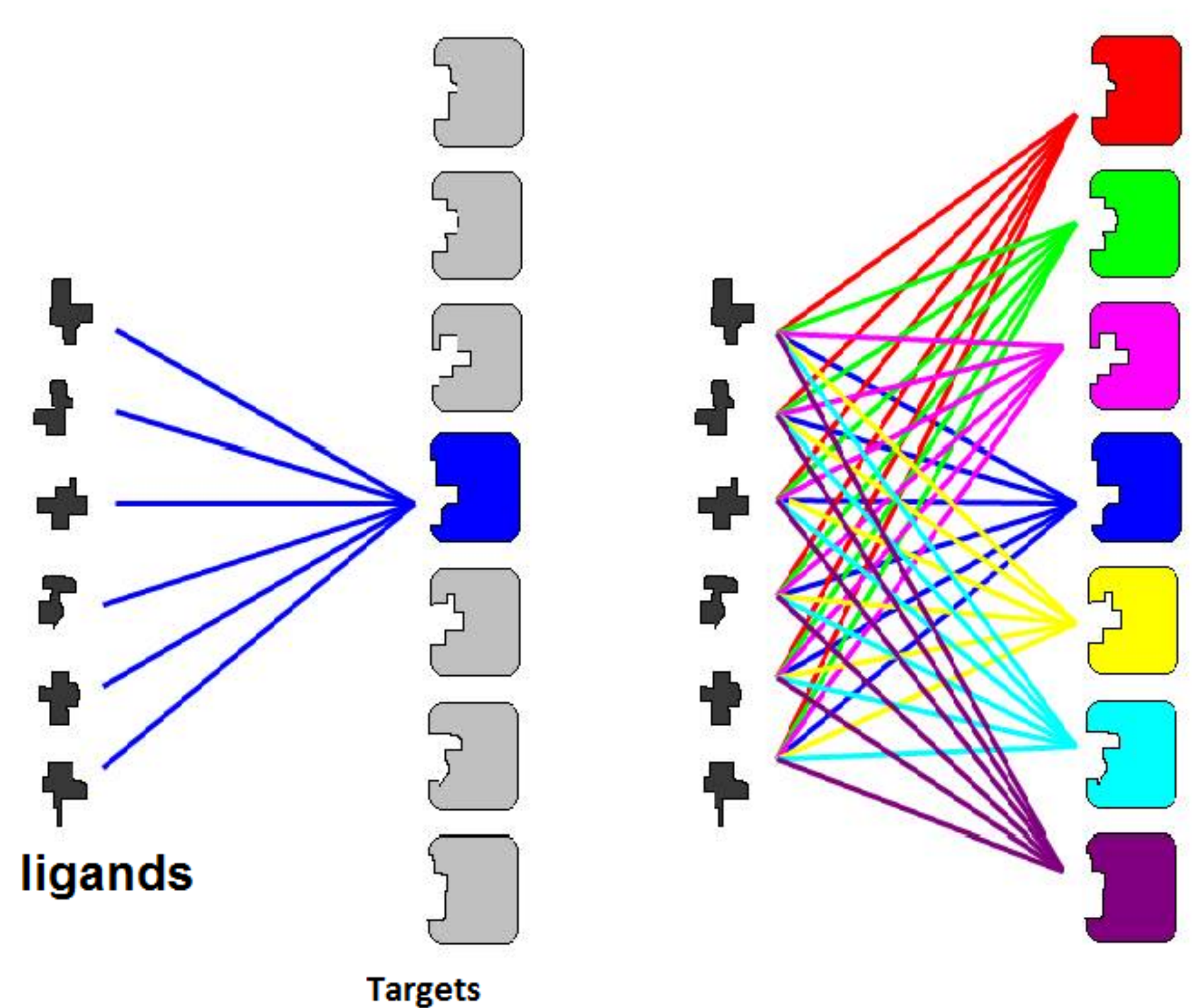
Проблема

События реакции лиганда с различными рецепторами не независимы. Классификатор, не учитывающий их, имеет неоптимальный результат. [2]

Задача

Необходимо построить

- вероятностную модель, учитывающую зависимости
- бинарный классификатор



Постановка задачи

Задана выборка $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\} = \mathcal{L} \sqcup \mathcal{T}$.
 $\mathbf{x}_i \in \mathbb{R}^n$. $\mathbf{y}_i \in \{0, 1, \square\}^l$ (задача класса MLC)
 \mathbf{X}, \mathbf{Y} — случайные величины

1. Восстановление плотности

$f(\mathbf{x}, \mathbf{y} | \mathbf{w}) = P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}; \mathbf{w})$ — модель классификации.
Максимизируется правдоподобие выборки:

$$\mathbf{w}^* = \arg \max_{\mathbf{w} \in \mathbf{W}} \ln Q(f | \mathbf{w}, \mathcal{D})$$

2. Бинарный классификатор

$L(\mathbf{y}, \mathbf{y}')$ — функция потерь

$$h(x) = \arg \min_{\mathbf{y} \in \mathbf{Y}} \mathbb{E}_{\mathbf{Y} | \mathbf{X} = \mathbf{x}} L(\mathbf{Y}, \mathbf{y})$$

Восстановление плотности

Probabilistic Classifier Chains [3]

- Выразим искомую величину $P(\mathbf{y} | \mathbf{x})$:

$$P(\mathbf{y} | \mathbf{x}) = P(y_1 | \mathbf{x}) \prod_{i=2}^l P(y_i | y_1, \dots, y_{i-1}, \mathbf{x})$$

- Задача распадается на n задач поиска

$$P(y_1 | \mathbf{x}), P(y_2 | y_1, \mathbf{x}), \dots, P(y_l | y_1, \dots, y_{l-1}, \mathbf{x})$$

- Каждую оцениваем при помощи логистической регрессии.
- Признаки для i -й: \mathbf{x} , а также y_1, \dots, y_{i-1}

Бинарный классификатор

Байесовское решающее правило:

$$h(x) = \arg \min_{\mathbf{y} \in \mathbf{Y}} \mathbb{E}_{\mathbf{Y} | \mathbf{X} = \mathbf{x}} L(\mathbf{Y}, \mathbf{y})$$

Все зависит от $L(\mathbf{y}, \mathbf{y}')$. Какая лучше? [4]

- Hamming Loss: $L(\mathbf{y}, \mathbf{y}') = \sum_{i=1}^l [y_i \neq y'_i]$.
 $h(x)$ не учитывает зависимости!
- Subset Loss: $L(\mathbf{y}, \mathbf{y}') = [\mathbf{y} \neq \mathbf{y}']$
- $L(\mathbf{y}, \mathbf{y}') = q(\sum_{i=1}^l [y_i \neq y'_i])$

Решения для разных существенно различны.

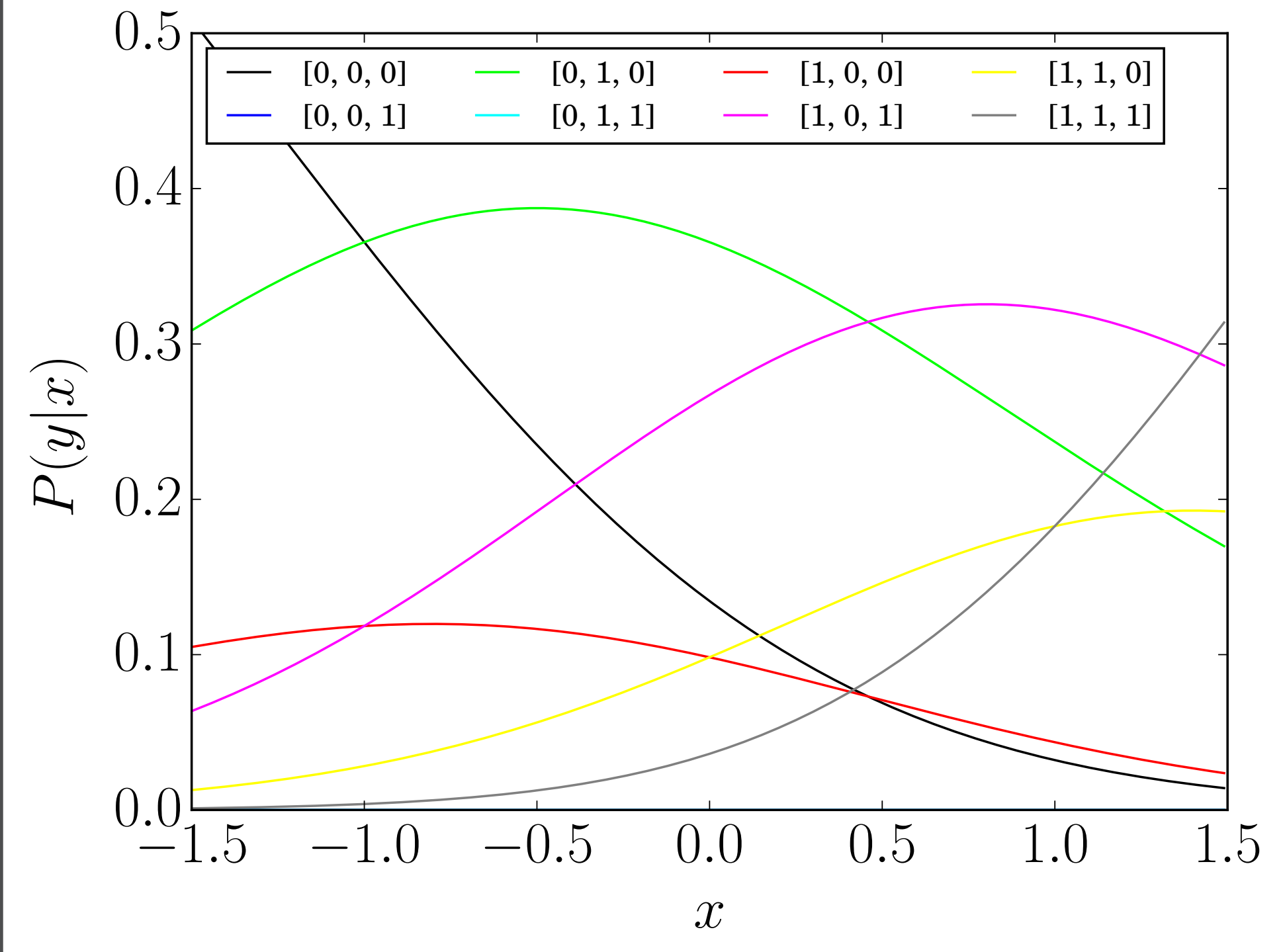
y_1	y_2	y_3	y_4	$P(\mathbf{y})$
0	0	0	0	0.30
0	1	1	1	0.17
1	0	1	1	0.18
1	1	0	1	0.17
1	1	1	0	0.18

Лучший по Subset Loss: (0, 0, 0, 0)

Лучший по Hamming Loss: (1, 1, 1, 1)

Модельные данные

1 признак, $l = 3$ класса. Плотность:



Модели:

- PCC + решающее правило
- Binary Relevance — 1 лог. регрессий

Результаты:

- Есть улучшение по Subset Loss
- Нет улучшений по Hamming Loss
- Нет улучшений по метрикам отдельных классов

Реальные данные

Взаимодействие лиганд и рецепторов. Признаки сгенерированы программой биохимической симуляции, ответы — результаты экспериментов.

- 165 признаков, 8000 объектов
- 12 классов (рецепторов), используется 3.
- Высокая мультиколлинеарность

В ответах имеется большое количество пропусков.

Результаты:

- Небольшое улучшение по Subset Loss
- Нет улучшений по Hamming Loss
- Нет улучшений по метрикам отдельных классов

Результаты

- Предложена модель для предсказания взаимодействия, учитывающая зависимости между классами
- Проведено сравнение модели с базовой
- PCC лучше BR по метрике Subset Loss
- Нет улучшений по отдельным классам

Дальнейшее развитие

- Улучшение показателей по классам
- Замена логистической регрессии на лучший алгоритм
- Вычисление для всей выборки

Список литературы

[1] Tong Q Xie XQ Myint KZ, Wang L. Molecular fingerprint-based artificial neural networks qsar for ligand biological activity predictions. *Molecular Pharmaceutics*, 2012.

[2] M. Popova. Feature selection and multi-task prediction of biological activity for nuclear receptors. 11(1):111–112, 2015.

[3] Eyke H.O Krzysztof Dembczynski, Weiwei Cheng. Bayes optimal multilabel classification via probabilistic classifier chains. 2010.

[4] Krzysztof Dembczynski. Multi-label classification: Label dependence, loss minimization, and reduction algorithms, 2013.

Вычислительный эксперимент

Использованы модельные данные, BR, PCC с различными $L(\mathbf{y}, \mathbf{y}')$

Метрика	BR	PCC (H)	PCC (M)	PCC (S)
Hamming	0.37 ± 0.009	0.36 ± 0.02	0.36 ± 0.02	0.38 ± 0.04
Hamming 1	0.31 ± 0.03	0.31 ± 0.03	0.31 ± 0.02	0.31 ± 0.05
Hamming 2	0.45 ± 0.04	0.45 ± 0.04	0.45 ± 0.03	0.49 ± 0.05
Hamming 3	0.34 ± 0.03	0.3 ± 0.03	0.31 ± 0.04	0.34 ± 0.03
Precision 1	0.7 ± 0.06	0.7 ± 0.06	0.73 ± 0.05	0.64 ± 0.05
Precision 2	0.55 ± 0.04	0.51 ± 0.01	0.47 ± 0.04	0.46 ± 0.07
Precision 3	0.7 ± 0.06	0.56 ± 0.05	0.5 ± 0.1	0.66 ± 0.05
Recall 1	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.03	0.71 ± 0.05
Recall 2	0.52 ± 0.1	0.53 ± 0.1	0.54 ± 0.09	0.48 ± 0.05
Recall 3	0.48 ± 0.1	0.53 ± 0.06	0.52 ± 0.07	0.49 ± 0.09
Subset	0.78 ± 0.03	0.77 ± 0.05	0.77 ± 0.05	0.62 ± 0.06