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

Цель

Предсказание взаимодействия двух типов молекул: лиганд и рецепторов.

Проблема

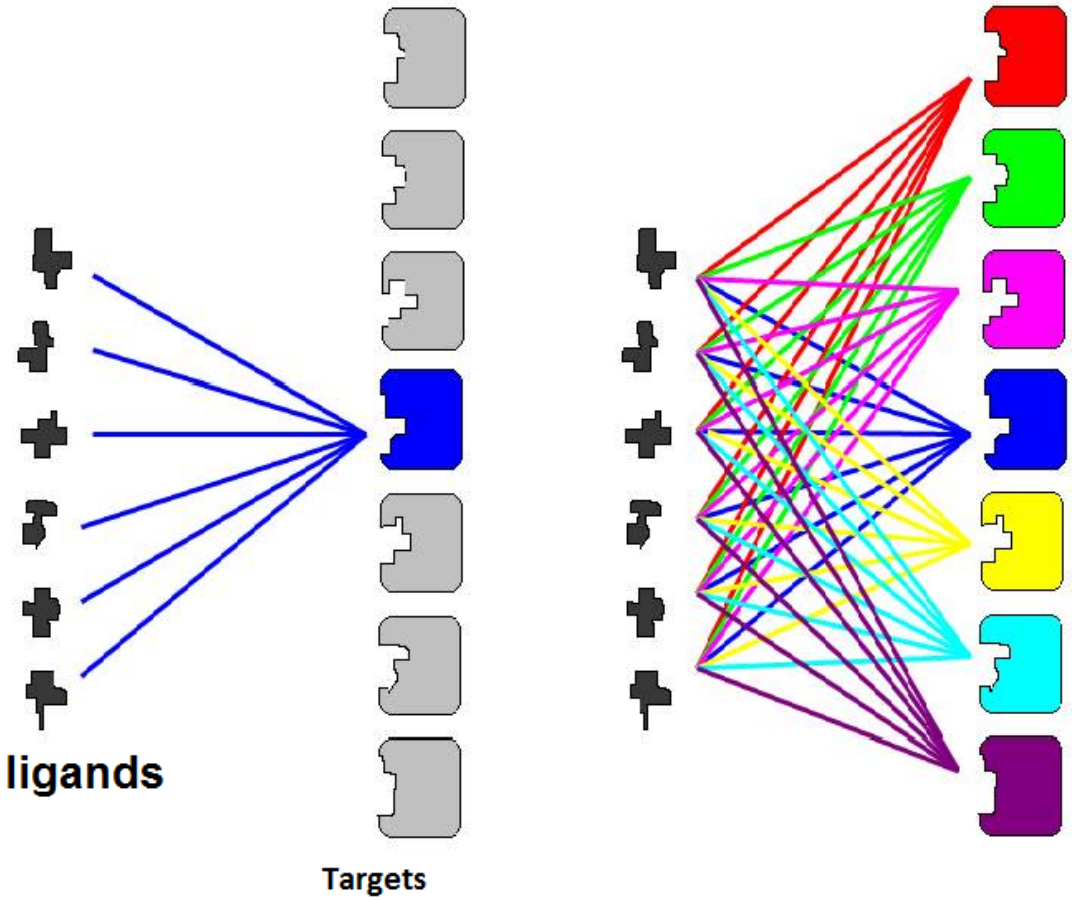
События реакции лиганда с различными рецепторами не независимы. Классификатор, не учитывающий их, имеет неоптимальный результат.

Задача

Необходимо построить

1. вероятностную модель, учитывающую зависимости
2. 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 $\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\} = \mathfrak{L} \sqcup \mathfrak{T}$. $\mathbf{x}_i \in \mathbb{R}^n$. $\mathbf{y}_i \in \{0, 1, \square\}^l$
$\mathbf{X}, \mathbf{Y}$ — случайные величины

1. Восстановление плотности

Модель классификации: функция $\mathrm{f} = P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}; \mathbf{w})$
Максимизируется правдоподобие выборки:

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

2. Бинарный классификатор

1. Функция потерь $L(\mathbf{y}, \mathbf{y}')$
2. $h(x) = \arg\min_{\mathbf{y} \in \mathbf{Y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} L(\mathbf{Y}, \mathbf{y})$

Модели сравниваются по различным метрикам с использованием кросс-валидации.

Решение. Вид модели f

Probabilistic Classifier Chains

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

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

2. Задача распадается на n задач поиска

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

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

Решение. 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?

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

Решения для разных существенно различны.

Решение. Бинарный классификатор

Иллюстрация проблемы

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

1. Лучший по Subset Loss: $(0, 0, 0, 0)$
2. Лучший по Hamming Loss: $(1, 1, 1, 1)$

Krzysztof Dembczynski, Multi-Label Classification: Label Dependence, Loss Minimization, and Reduction Algorithms, 2013

Вычислительный эксперимент

Цель эксперимента

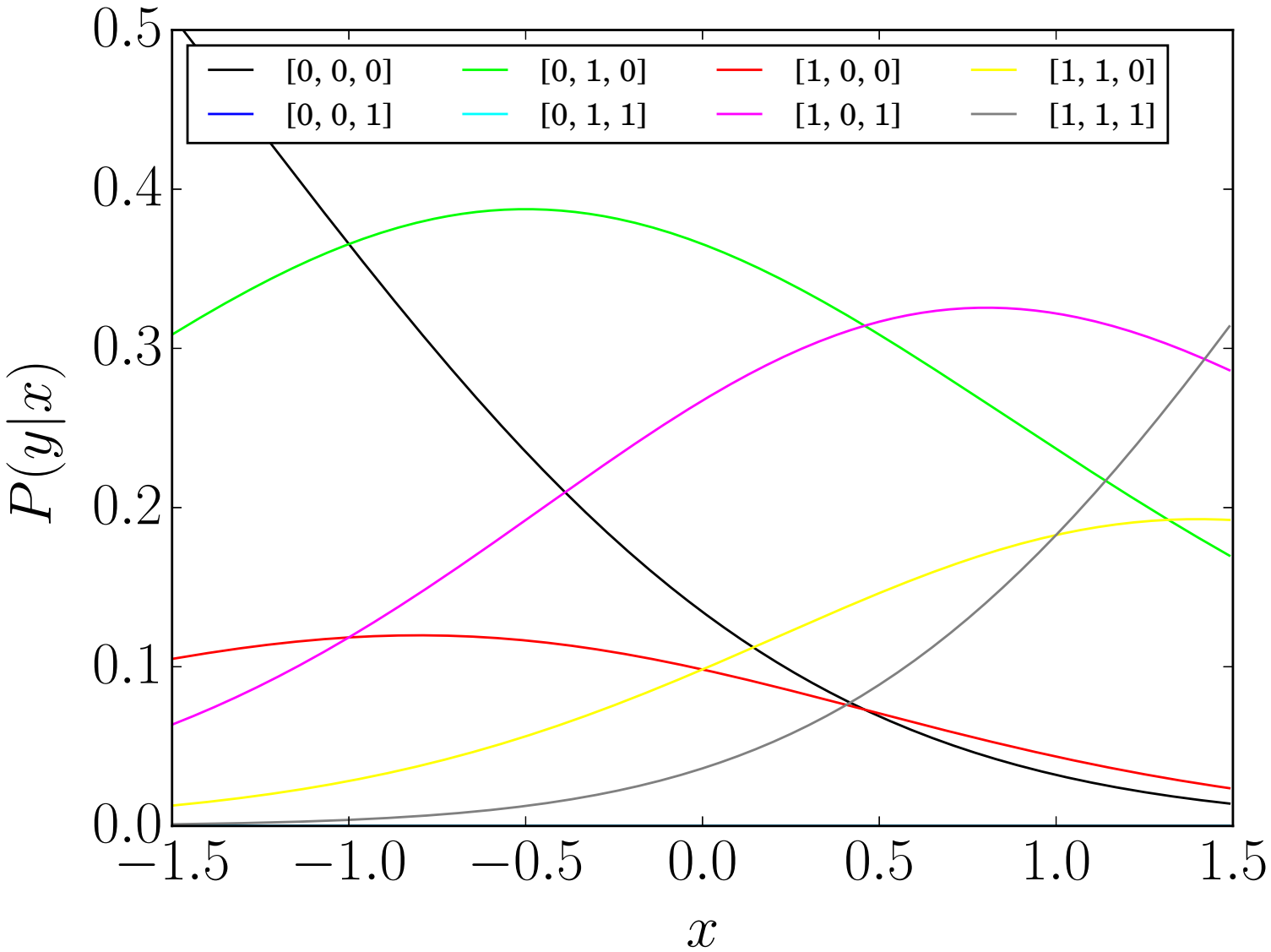
1. Сравнение бинарных классификаторов
2. Подбор гиперпараметров
3. Характеристики алгоритма

Сравниваемые модели

1. Binary Relevance — без зависимостей
2. **PCC** — предлагаемое решение

Вычислительный эксперимент. Модельные данные

1 признак, $l = 3$ класса. Плотность:



Вычислительный эксперимент. Модельные данные

Метрика	BR	PCC (H)	PCC (M)	PCC (S)
AUC 1	0.69 ± 0.03	0.69 ± 0.03	0.69 ± 0.02	0.69 ± 0.05
AUC 2	0.55 ± 0.04	0.55 ± 0.04	0.56 ± 0.03	0.51 ± 0.04
AUC 3	0.65 ± 0.04	0.66 ± 0.02	0.64 ± 0.04	0.64 ± 0.04
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

Есть улучшение только по Subset Loss.

Вычислительный эксперимент. Реальные данные

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

1. 165 признаков, 8000 объектов
2. 12 классов (рецепторов), используется 3.
3. Высокая мультиколлинеарность

В ответах имеется большое количество пропусков.

Вычислительный эксперимент. Реальные данные

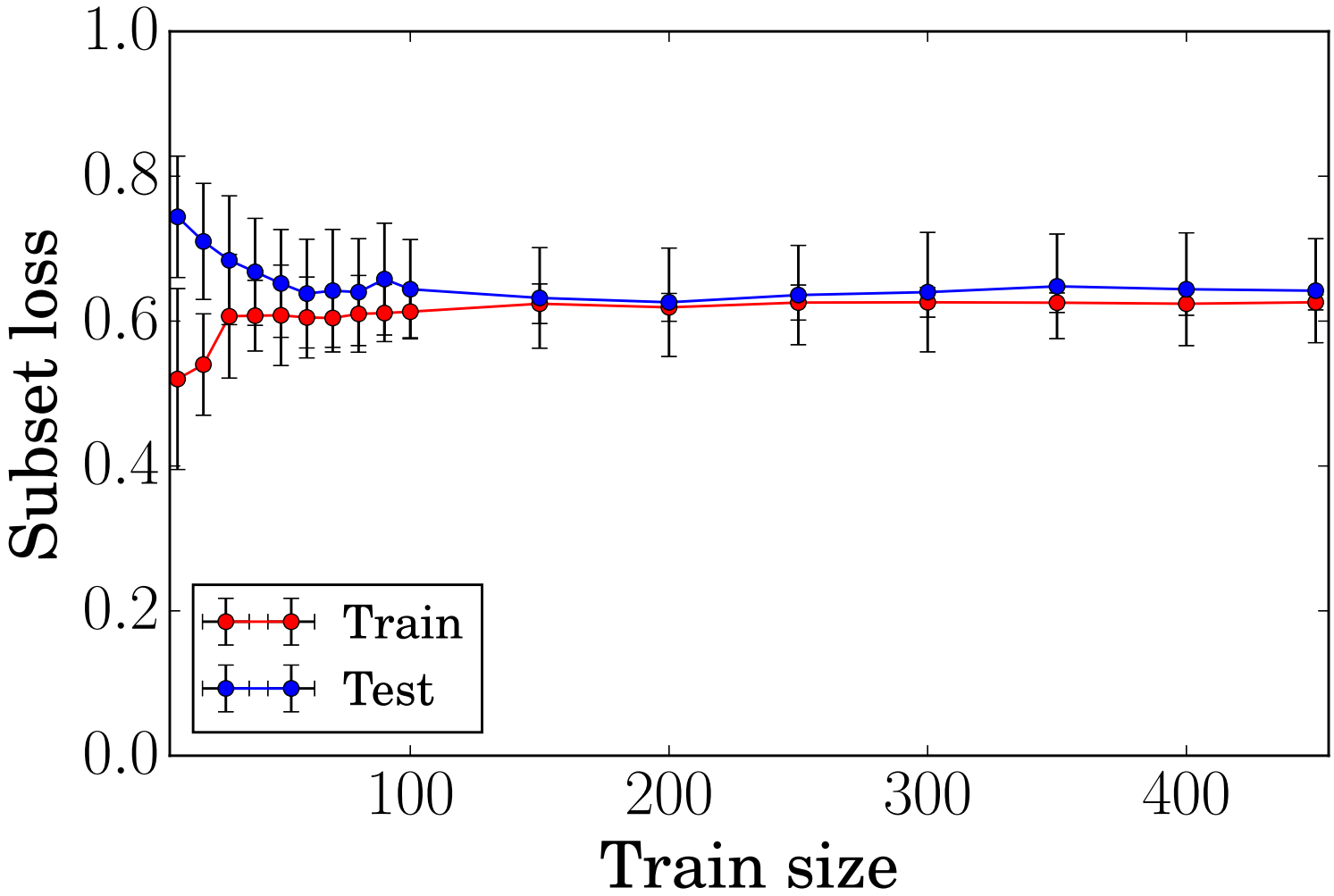
1,2,3 = NR-AhR, NR-AR-LBD, NR-Aromatase

Метрика	BR	PCC (H)	PCC (M)	PCC (S)
AUC 1	0.58 ± 0.03	0.58 ± 0.03	0.57 ± 0.02	0.58 ± 0.02
AUC 2	0.61 ± 0.06	0.61 ± 0.06	0.62 ± 0.06	0.61 ± 0.05
AUC 3	0.55 ± 0.01	0.54 ± 0.01	0.53 ± 0.01	0.54 ± 0.01
Hamming	0.15 ± 0.01	0.17 ± 0.01	0.19 ± 0.02	0.17 ± 0.02
H 1	0.21 ± 0.03	0.21 ± 0.03	0.24 ± 0.02	0.21 ± 0.03
H 2	0.045 ± 0.01	0.041 ± 0.007	0.041 ± 0.008	0.041 ± 0.006
H 3	0.2 ± 0.02	0.25 ± 0.01	0.29 ± 0.03	0.25 ± 0.03
Precision 1	0.79 ± 0.1	0.79 ± 0.1	0.79 ± 0.1	0.82 ± 0.1
Precision 2	0.91 ± 0.1	0.88 ± 0.1	0.91 ± 0.1	0.88 ± 0.1
Precision 3	0.76 ± 0.07	0.82 ± 0.09	0.78 ± 0.09	0.82 ± 0.08
Recall 1	0.17 ± 0.06	0.17 ± 0.06	0.15 ± 0.04	0.18 ± 0.05
Recall 2	0.22 ± 0.1	0.23 ± 0.1	0.24 ± 0.1	0.23 ± 0.1
Recall 3	0.1 ± 0.02	0.085 ± 0.02	0.071 ± 0.02	0.086 ± 0.02
Subset	0.32 ± 0.02	0.34 ± 0.02	0.46 ± 0.03	0.3 ± 0.03

Небольшое улучшение по Subset Loss

Вычислительный эксперимент. Размер выборки

Модельные данные.



150 объектов достаточно.

Заключение

Результаты

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

Планы на будущее

1. Улучшение показателей по классам
2. Замена логистической регрессии на лучший алгоритм
3. Вычисление для всей выборки