

Вероятностный подход для задачи предсказания биологической активности ядерных рецепторов*

Володин С. Е., Попова М., Стрижов В. В.

sergei.volodin@phystech.edu

Московский физико-технический институт

Решается задача предсказания биологической активности молекул протеинов (лиганд) с рецепторами: по признакам лиганда необходимо оценить вероятность связывания этой молекулы с одним или несколькими клеточными рецепторами и построить бинарный классификатор. Экспертные знания в области биохимии и фармакологии дают основания предполагать, что факты связывания одних и тех же молекул с различными рецепторами не независимы. В данной работе предлагается модель, позволяющая строить предсказания сразу для группы рецепторов, учитывая их схожесть. Модель оценивает условные вероятности принадлежности классам. В работе проводится вычислительный эксперимент на реальных данных, в ходе которого предложенная модель сравнивается с независимыми моделями в терминах нескольких функционалов качества.

Ключевые слова: классификация, вероятность, classifier chains, multi-label, логистическая регрессия, PCC.

Probabilistic prediction of nuclear receptors' biological activity*

Sergey Volodin, Maria Popova, Vadim Strijov

The paper solves the problem of prediction of biological activity of proteins (ligands) with cell receptors. Given the ligand feature description, the probability of reaction with one or several cell receptors must be estimated alongside with constructing a binary classifier. Expert knowledge in biochemistry and pharmacology state that events of reaction with different cell receptors are not independent. This paper suggests a model allowing to make predictions for groups of receptors exploiting their similarities. The model evaluates conditional probabilities of classes. A computational experiment on real data is conducted in which the proposed model is compared with simple models in terms of several metrics.

Keywords: classification, probability, classifier chains, multi-label, logistic regression, PCC.

Введение

Проблема предсказания биологической активности лигандов и рецепторов является актуальной задачей в области биохимии и фармакологии [1, 2, 3, 4, 5, 6]. Данная статья посвящена решению этой задачи методами машинного обучения.

Компьютерное моделирование взаимодействия молекул является распространенным методом предсказания биологической активности клеточных рецепторов [4, 1]. Однако такой способ требует знания точной структуры лиганд, которая не всегда известна. По этой причине развитие методов машинного обучения [7], позволяющих делать предсказания на основании только числовых признаков лиганд, является актуальным.

Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Эксперт И. О. Консультант: Попова М.

Существует два основных подхода к решению описанной задачи. В рамках первого из них для каждого клеточного рецептора строятся независимые модели. Так, например в [8, 5] применяется метод опорных векторов, в [2, 3] — нейронные сети, а в [9] — метод к ближайших соседей. Второй подход подразумевает построение одной модели для предсказания активности группы рецепторов. Такой подход позволяет строить более сложные модели, учитывающие информацию о схожести рецепторов [6]. В [10] проведен сравнительный анализ обоих подходов.

Таким образом, данная задача решается многими способами. Тем не менее, как показывает сопоставление результатов [10], лучшим оказывается второй подход, т.е. классификаторы, учитывающие при обучении все рецепторы сразу, а не независимо друг от друга. В данном случае это означает использование нескольких классификаторов и объединение их в «цепочку» [11, 12, 13]. Как показывает практика, обучение нескольким задачам сразу дает существенный прирост в качестве конечного алгоритма по сравнению с рассмотрением этих задач по-отдельности [14, 15, 13].

В данной работе предлагается усовершенствованный метод classifier chains [13] — вероятностная модель последовательного вывода для предсказания биологической активности рецепторов [16, 14]. Предложенный алгоритм относится ко второму подходу, то есть позволяет строить предсказания для группы рецепторов, а также допускает добавление новых без необходимости повторного обучения. Проведен вычислительный эксперимент на реальных данных, в котором набор независимых моделей сравнивался с моделью последовательного вывода. Построенные модели сравнивались по нескольким критериям качества.

Постановка задачи классификации

Задана выборка $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{L}}$, $\mathcal{L} = \{1, \dots, m\}$ — m пар объект-ответ. Каждый из объектов $\mathbf{x}_i \in \mathbb{R}^n$ — вектор действительных чисел. Объект может принадлежать каждому из l , что представляется вектором ответов $\mathbf{y}_i \in \{0, 1, \square\}^l$, 1 означает принадлежность классу, а \square означает пропуск в данных. Выборка разбита на обучающую и контрольную: $\mathcal{D} = \mathcal{L} \sqcup \mathcal{T}$

Определяются \mathbf{X}, \mathbf{Y} — случайные величины. Считается, что между классами есть зависимости:

$$P(\mathbf{Y}|\mathbf{X}) \neq \prod_{j=1}^l P(y_j|\mathbf{X})$$

Моделью классификации называется функция

$$f: \mathbf{W} \times \mathbf{X} \times \mathbf{Y} \rightarrow [0, 1],$$

где \mathbf{W} — множество параметров, $\mathbf{w} \in \mathbf{W}$ — вектор параметров модели. Значение f — апостериорная вероятность ответов \mathbf{y} при фиксированном \mathbf{x} :

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

Функция потерь для значения параметра \mathbf{w} и подвыборки \mathcal{Z} определяется через функцию правдоподобия модельного распределения:

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

Требуется найти вектор параметров $\mathbf{w}^* \in \mathbf{W}$, минимизирующий Q на обучающей выборке \mathfrak{L} :

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

Для вывода бинарного классификатора из вероятностной модели $P(\mathbf{y}|\mathbf{x})$ вводится функция потерь, т.е. штраф за ответ \mathbf{y} при правильном ответе $\mathbf{y}' \in \mathbf{Y}$:

$$L: Y \times Y \rightarrow \mathbb{R}$$

Бинарный классификатор $\mathbf{h}: \mathbf{X} \rightarrow \mathbf{Y}$ получается [14] при помощи Байесовского решающего правила:

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

В качестве дополнительного критерия качества модели используются значения Subset Loss для векторов ответов, а также значения Hamming Loss и других метрик для каждого класса j на контрольной выборке \mathfrak{T} при 5 различных разбиениях.

Поскольку выборка содержит пропуски, разбиения должны быть построены таким образом, чтобы в каждой подвыборке было достаточное количество объектов с известным значением каждого признака.

Описание алгоритма

Таким образом, задача предсказания разбивается на два этапа:

1. Поиск параметра модели \mathbf{w} максимизацией правдоподобия выборки на семействе распределений $P(\mathbf{y}|\mathbf{x}; \mathbf{w})$. В результате решения задачи получается модель $P_{\mathbf{w}^*}(\mathbf{y}|\mathbf{x})$ как функция двух переменных
2. Поиск оптимального бинарного классификатора $h: \mathbf{X} \rightarrow \mathbf{Y}$, использующего найденное распределение $P(\mathbf{y}|\mathbf{x})$. Конкретная функция получается применением Байесовского решающего правила для каждого \mathbf{x} , подлежащего классификации. Конкретный классификатор зависит от выбранной функции потерь L .

Часть 1. Предлагаемый вид модели

Решим первую часть поставленной задачи, используя метод, описанный в [14].

Рассмотрим искомую величину

$$P(\mathbf{y}|\mathbf{x})$$

Докажем равенство

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

Рассмотрим величину

$$P(y_i|y_1, \dots, y_{i-1}, \mathbf{x}) = \frac{P(y_1, \dots, y_i, \mathbf{x})}{P(y_1, \dots, y_{i-1}, \mathbf{x})}$$

Подставим их в произведение, получим телескопическое произведение:

$$P(y_1|\mathbf{x}) \prod_{i=2}^l P(y_i|y_1, \dots, y_{i-1}, \mathbf{x}) = \frac{P(y_1, \mathbf{x})}{P(\mathbf{x})} \frac{P(y_1, y_2, \mathbf{x})}{P(y_1, \mathbf{x})} \cdot \dots \cdot \frac{P(y_1, \dots, y_l, \mathbf{x})}{P(y_1, \dots, y_{l-1}, \mathbf{x})} = P(\mathbf{y}|\mathbf{x}) \blacksquare$$

Таким образом, для моделирования вероятности $P(\mathbf{y}|\mathbf{x})$ можно использовать условные вероятности классов

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

Каждую из l этих вероятностей будем оценивать при помощи логистической регрессии. Обозначим

$$(x)_y = \begin{cases} x, & y = 1 \\ 1 - x & y = 0 \end{cases}$$

Обозначим

$$g_i(y_1, \dots, y_{i-1}, \mathbf{x}) = P(y_i = 1|y_1, \dots, y_{i-1}, \mathbf{x})$$

Получаем выражение вероятности $P(\mathbf{y}|\mathbf{x})$ через функции g_i :

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

Вероятности

$$P(y_i = 1|y_1, \dots, y_{i-1}, \mathbf{x}) = g_i(y_1, \dots, y_{i-1}, \mathbf{x})$$

предсказываются при помощи логистической регрессии, т.е.

$$g_i(y_1, \dots, y_{i-1}, \mathbf{x}) = \sigma(\mathbf{w}_i^T \|y_1 \dots y_{i-1} \mathbf{x}^T\|^T + w_i^0), \text{ где } \sigma(x) = \frac{1}{1 + e^{-x}}$$

Получаем семейство моделей

$$P(\mathbf{y}|\mathbf{x}) = (\sigma(\mathbf{w}_1^T \mathbf{x} + w_1^0))_{y_1} \prod_{i=2}^l (\sigma(\mathbf{w}_i^T \|y_1 \dots y_{i-1} \mathbf{x}^T\|^T + w_i^0))_{y_i}$$

Таким образом, общая задача оптимизации \mathbf{w}^* распадается на l независимых оптимизационных задач максимизации правдоподобия, т.е. на обучение l логистических регрессий. i -я логистическая регрессия принимает в качестве признаков \mathbf{x} , а также ответы y_1, \dots, y_{i-1}

Данный алгоритм называется PCC (Probabilistic Classifier Chain) [14]

Часть 2. Бинарный классификатор

Решим вторую часть задачи, т.е. построим бинарный классификатор по известному распределению $P(\mathbf{y}|\mathbf{x})$, выбирая некоторую функцию потерь (см. [14]).

При фиксированной функции потерь L и объекте $\mathbf{x} \in \mathbf{X}$ оптимальное предсказание $\mathbf{h}(\mathbf{x}) \in \mathbf{Y}$ в соответствии с Байесовским решающим правилом имеет вид [14]:

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

В качестве примеров рассмотрим следующие функции потерь $L(\mathbf{y}, \mathbf{y}')$ и приведем полученный алгоритм $h(\mathbf{x})$ [14]:

1. Hamming Loss. Получаем $h_i(\mathbf{x}) = \text{sign}(P(y_i = 1|\mathbf{x}) - \frac{1}{2})$
2. Subset 0/1 Loss. Получаем $h(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathbf{Y}} P(\mathbf{y}|\mathbf{x})$
3. Rank Loss. Получаем $f_i(\mathbf{x}) = P(y_i = 1|\mathbf{x})$

Используемая вероятность $P(y_i = 1|\mathbf{x})$ может быть получена из известного распределения $P(\mathbf{y}|\mathbf{x})$ по формуле полной вероятности:

$$P(y_i = 1|\mathbf{x}) = \sum_{\mathbf{y} \in \{0,1\}^l} [y_i = 1] P(\mathbf{y}|\mathbf{x})$$

Таким образом, искомые вероятности выражаются через известное распределение $P(\mathbf{y}|\mathbf{x})$.

Часть 3. Работа с пропусками

Приведенный выше алгоритм РСС построения $P(\mathbf{y}|\mathbf{x})$ по имеющейся обучающей выборке неприменим для выборок, для которых в ответах могут содержаться пропуски: $y_i \in \{0, 1, \square\}$. Эта проблема решается следующим образом:

1. Логистические регрессии $1, \dots, l$ обучаются последовательно
2. Для обучения i -й логистической регрессии берутся объекты с известным значением признака y_i
3. Предыдущие неизвестные значения признаков y_1, \dots, y_{i-1} предсказываются частично уже построенным РСС для классов $1, \dots, i-1$.

Алгоритмы

Алгоритм 1 Обучение РСС для выборок без пропусков

Вход: Обучающая выборка $\mathcal{L} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in L}$

Выход: Векторы $\mathbf{w}_i \in \mathbb{R}^{n+i-1}$, $i \in \overline{1, l}$

- 1: $u_j \leftarrow j$ -й столбец матрицы y_{ij} , $j \in \overline{1, l}$
 - 2: для $i = 1, \dots, l$
 - 3: $X^i \leftarrow \|X y_1 \dots y_{i-1}\|^\square$. Эта матрица имеет строки X_j^i
 - 4: $\mathbf{w}_i = \arg \max_{j \in L} \prod (\sigma(\mathbf{w}_i^T X_j^i))_{y_{ij}}$ — обучение логистической регрессии
- вернуть $\mathbf{w}_1, \dots, \mathbf{w}_l$
-

Алгоритм 2 Предсказание вероятности $P(\mathbf{y}|\mathbf{x})$ для пары объект-ответ

Вход: Объект $\mathbf{x} \in \mathbb{R}^n$, векторы \mathbf{w}_i , пороги w_i^0 , вектор $\mathbf{y} \in \{0, 1\}^m$

Выход: Условная вероятность $P(\mathbf{y}|\mathbf{x}) \in [0, 1]$

- 1: $P \leftarrow 1$
 - 2: для $i = 1, \dots, l$
 - 3: $\mathbf{x}' \leftarrow \|\mathbf{x}^T y_1 \dots y_{i-1}\|^\square$
 - 4: $P \leftarrow P \cdot (\sigma(\mathbf{w}_i^T \mathbf{x}' + w_i^0))_{y_i}$
- вернуть P
-

Вычислительный эксперимент

Целью эксперимента является получение характеристик предложенного алгоритма и сравнение результатов с базовым алгоритмом. Также в ходе эксперимента находятся значения гиперпараметров исходя из оптимизации функций ошибок.

Базовый алгоритм использует подход Binary Relevance [14], в котором зависимости между классами не учитываются. Таким образом, алгоритм представляет собой l независимых логистических регрессий, по одному классификатору для каждого класса. Предлагаемый алгоритм, РСС, учитывает зависимости между классами.

Для решения второй части задачи в предлагаемом алгоритме рассматриваются следующие функции потерь:

1. Subset 0/1 loss: $L(\mathbf{y}, \mathbf{y}') = [\mathbf{y} \neq \mathbf{y}']$
2. Hamming loss: $L(\mathbf{y}, \mathbf{y}') = \sum_{i=1}^l [y_i \neq y'_i]$
3. Функционал $L(\mathbf{y}, \mathbf{y}') = q\left(\sum_{i=1}^l [y_i \neq y'_i]\right)$, где $q(t)$ задана в натуральных точках $t \in \overline{0, l}$ и подлежит оптимизации.

Для полученных результатов бинарных классификаторов также сравниваются значения Precision, Recall, Hamming loss и AUC для каждого класса, а также Hamming Loss и Subset Loss в целом. Для оценки стандартного отклонения используется 5-fold разбиение. Эксперимент проведен на модельных и реальных данных.

Модельные данные

Используется следующая вероятностная модель для генерации выборки:

Выборка $\mathfrak{D} = \{(x_i, \mathbf{y}_i)\}_{i \in \mathcal{L}}$, $\mathcal{L} = \{1, \dots, m\}$ — m пар объект-ответ. Каждый из объектов $x_i \in [-1.5, 1.5]$ — действительное число. Объект может принадлежать каждому из $l = 3$ классов, что представляется вектором ответов $\mathbf{y}_i \in \{0, 1\}^l$, 1 означает принадлежность классу. В модельных данных пропуски в ответах отсутствуют.

Вероятность принадлежности объекта x к классам $\mathbf{y} \in \{0, 1\}^3$ $P(\mathbf{y}|x)$ задается по формуле [14]:

$$P(y_1, y_2, y_3|x) = (f_1(x))_{y_1} (f_2(x, y_1))_{y_2} (f_3(x, y_1, y_2))_{y_3},$$

где f_1, f_2, f_3 заданы следующим образом:

$$\begin{aligned} f_1(x) &= \sigma(x) \\ f_2(x, y_1) &= \sigma(x - 2y_1 + 1) \\ f_3(x, y_1, y_2) &= \sigma(x + 12y_1 - 2y_2 - 11) \end{aligned}$$

Выборка содержит 500 объектов. Генерация производится следующим образом:

1. Выбирается $x \sim u[-1.5, 1.5]$ — из равномерного распределения
2. Выбирается \mathbf{y} для данного x в соответствии с формулой.

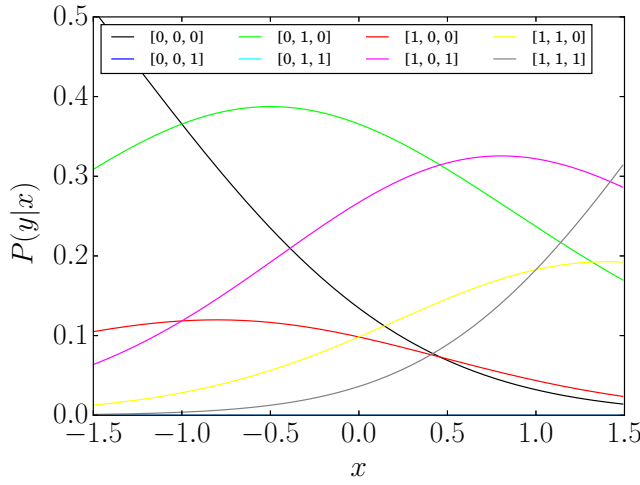
Полученные плотности $P(\mathbf{y}|x)$ изображены на графике 1.

Для сравнения алгоритмов использовались следующие метрики: AUC_i — AUC для каждого класса, H_i — Hamming Loss для каждого класса, P_i — Precision, R_i — Recall, S — Subset Loss, H — общий Hamming Loss.

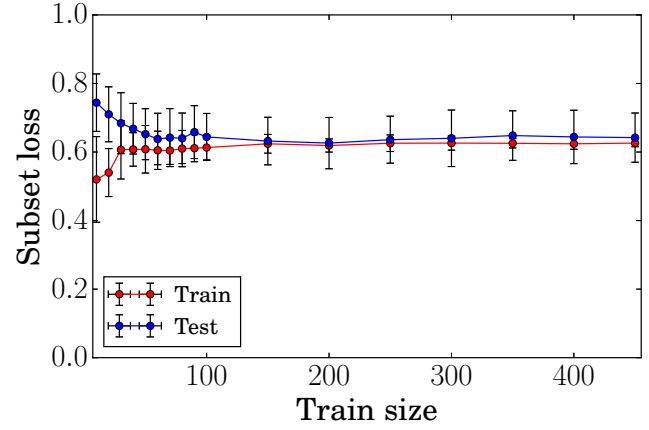
Для контроля переобучения используется 5-fold кросс-валидация.

В качестве функций потерь для РСС использовались следующие: H (Hamming Loss), S (Subset Loss), а также M — функция вида

$$L(\mathbf{y}, \mathbf{y}') = q\left(\sum_{i=1}^l [y_i \neq y'_i]\right),$$



(a) Плотность модельных данных



(b) Зависимость ошибки на обучении и контроле от размера обучающей выборки

Рис. 1. Модельные данные

Функция q определена в точках $\overline{0, l} = \overline{0, 3}$. Проведена оптимизация q по различным метрикам итогового алгоритма. Значения q в точках $(0, 1, 2, 3)$ имеют вид $(0, a_1, a_2, 10)$, где a_1, a_2 подлежат перебору.

Оптимальная функция q зависит от метрики и класса, для которого вычисляется данная метрика.

Показано, что для оптимизации Subset Loss $a_1 = 10, a_2 = 10$, а для оптимизации суммарного Hamming Loss $a_1 = 2, a_2 = 5$. В качестве q_M взята последняя.

Результаты представлены в таблице 1. Наблюдается серьезное улучшение в Subset Loss для PCC (S). Остальные изменения в пределах погрешности.

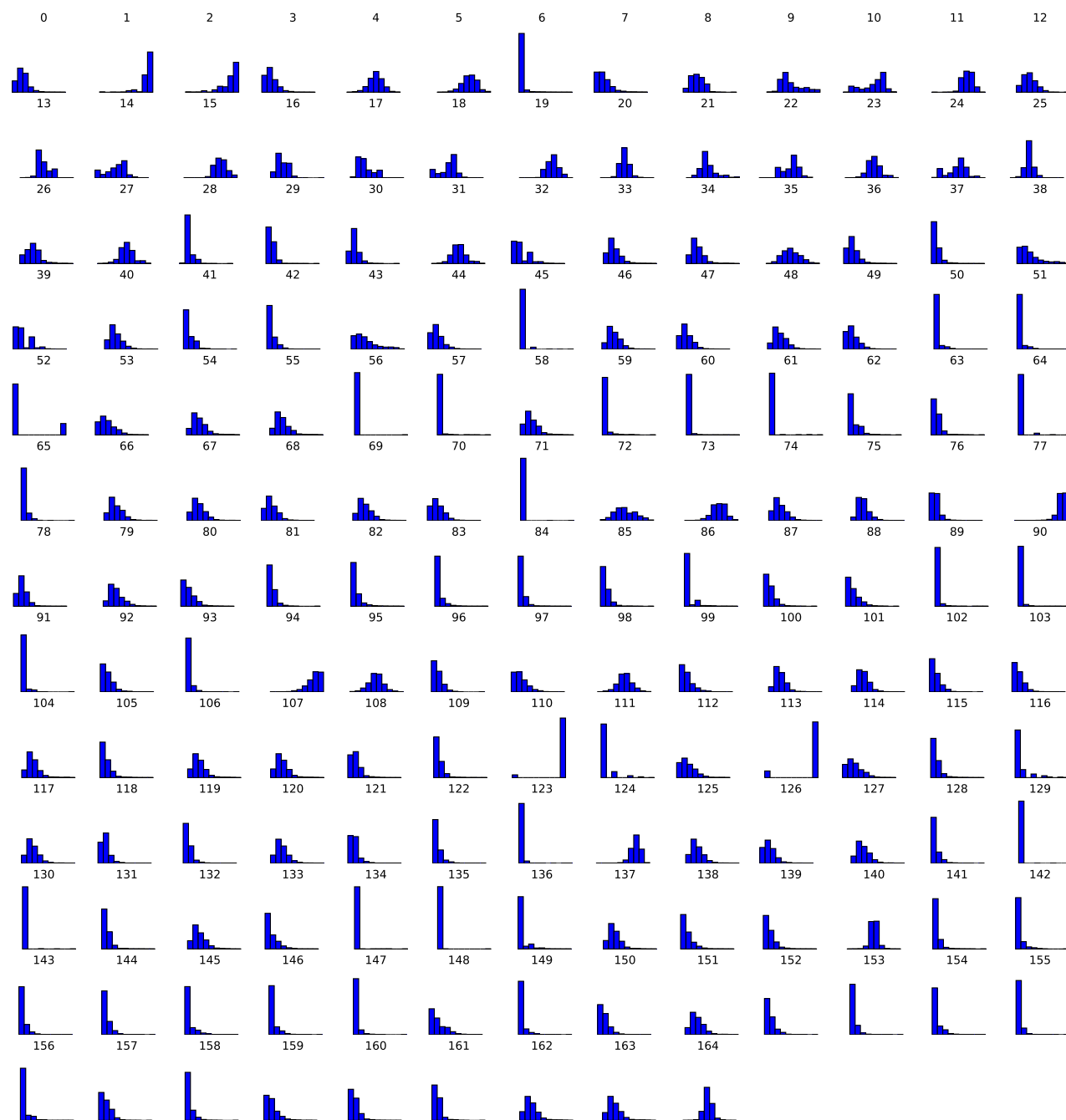
Таблица 1. Сравнение алгоритмов на модельных данных

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
H	0.37 ± 0.009	0.36 ± 0.02	0.36 ± 0.02	0.38 ± 0.04
H 1	0.31 ± 0.03	0.31 ± 0.03	0.31 ± 0.02	0.31 ± 0.05
H 2	0.45 ± 0.04	0.45 ± 0.04	0.45 ± 0.03	0.49 ± 0.05
H 3	0.34 ± 0.03	0.3 ± 0.03	0.31 ± 0.04	0.34 ± 0.03
P 1	0.7 ± 0.06	0.7 ± 0.06	0.73 ± 0.05	0.64 ± 0.05
P 2	0.55 ± 0.04	0.51 ± 0.01	0.47 ± 0.04	0.46 ± 0.07
P 3	0.7 ± 0.06	0.56 ± 0.05	0.5 ± 0.1	0.66 ± 0.05
R 1	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.03	0.71 ± 0.05
R 2	0.52 ± 0.1	0.53 ± 0.1	0.54 ± 0.09	0.48 ± 0.05
R 3	0.48 ± 0.1	0.53 ± 0.06	0.52 ± 0.07	0.49 ± 0.09
S	0.78 ± 0.03	0.77 ± 0.05	0.77 ± 0.05	0.62 ± 0.06

График 1 показывает зависимость функции ошибки Subset Loss на обучающей и контрольной выборке от мощности обучающей выборки. Видно, что при $|\mathcal{L}| < 100$ ошибка на контроле сильно больше ошибки на обучении, т.е. возникает переобучение. При $|\mathcal{L}| \gtrsim 150$ этот эффект уходит, и ошибки становятся примерно равны.

Реальные данные

Эксперимент проведен на реальных данных, имеющих двойное происхождение. Объектами являются лиганды, их признаки \mathbf{x}_i смоделированы при помощи специальной программы. Ответы $\mathbf{y}_i = (y_{i1}, \dots, y_{il})$ являются результатами биохимических экспериментов, показывающих, связывается ли данный лиганд с рецептором j . Пропуск в ответах означает, что эксперимент либо не был проведен, либо не позволяет с достаточной уверенностью говорить о каком-либо результате. Каждый объект имеет 165 признаков. Признаки являются химическими параметрами молекулы. В выборке содержится 8513 объектов, количество объектов с измеренным ответом j составляет около половины. В таблице 2 указано точное распределение ответов по классам. График 2 показывает распределение объектов по значениям всех 165 признаков. Видно, что большинство распределений унимодальные.



(а) Распределение объектов по значениям признаков для реальных данных

Рис. 2. Распределение объектов по признакам

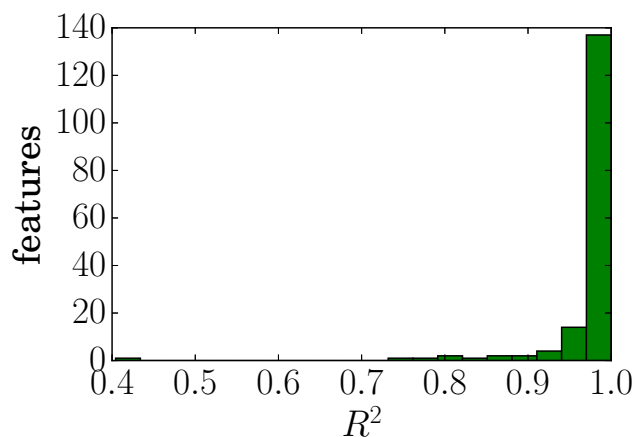
(a) Гистограмма R^2 для реальных данных**Рис. 3.** Реальные данные

График 3 показывает распределение признаков по значению $R^2 = 1 - \frac{1}{\text{VIF}}$. Видно, что данные обладают высокой мультиколлинеарностью (большинство признаков имеют R^2 , близкий к 1)

Таблица 2. Количество связывающихся с рецепторами лигандов

Рецептор	Неизвестно	Не связывается	Связывается
NR-AhR	3413 (40%)	4503 (52%)	597 (7%)
NR-AR-LBD	3213 (37%)	5129 (60%)	171 (2%)
NR-AR	2904 (34%)	5398 (63%)	211 (2%)
SR-MMP	3925 (46%)	3870 (45%)	718 (8%)
NR-ER	3746 (44%)	4232 (49%)	535 (6%)
SR-HSE	3309 (38%)	4961 (58%)	243 (2%)
SR-p53	3174 (37%)	5029 (59%)	310 (3%)
NR-PPAR-gamma	3393 (39%)	4987 (58%)	133 (1%)
SR-ARE	3791 (44%)	4029 (47%)	693 (8%)
NR-Aromatase	4544 (53%)	3835 (45%)	134 (1%)
SR-ATAD5	2951 (34%)	5360 (62%)	202 (2%)
NR-ER-LBD	3107 (36%)	5168 (60%)	238 (2%)

Таблица 3. Значение AUC для различных рецепторов и моделей классификации

Рецептор	Binary Relevance	Random Forest [17]
NR-AhR	0.83 ± 0.03	0.93
NR-AR-LBD	0.86 ± 0.08	0.88
NR-AR	0.83 ± 0.09	0.83
SR-MMP	0.87 ± 0.03	0.95
NR-ER	0.78 ± 0.04	0.81
SR-HSE	0.79 ± 0.04	0.86
SR-p53	0.79 ± 0.07	0.88
NR-PPAR-gamma	0.79 ± 0.04	0.86
SR-ARE	0.78 ± 0.02	0.84
NR-Aromatase	0.81 ± 0.05	0.84
SR-ATAD5	0.81 ± 0.06	0.83
NR-ER-LBD	0.80 ± 0.07	0.83

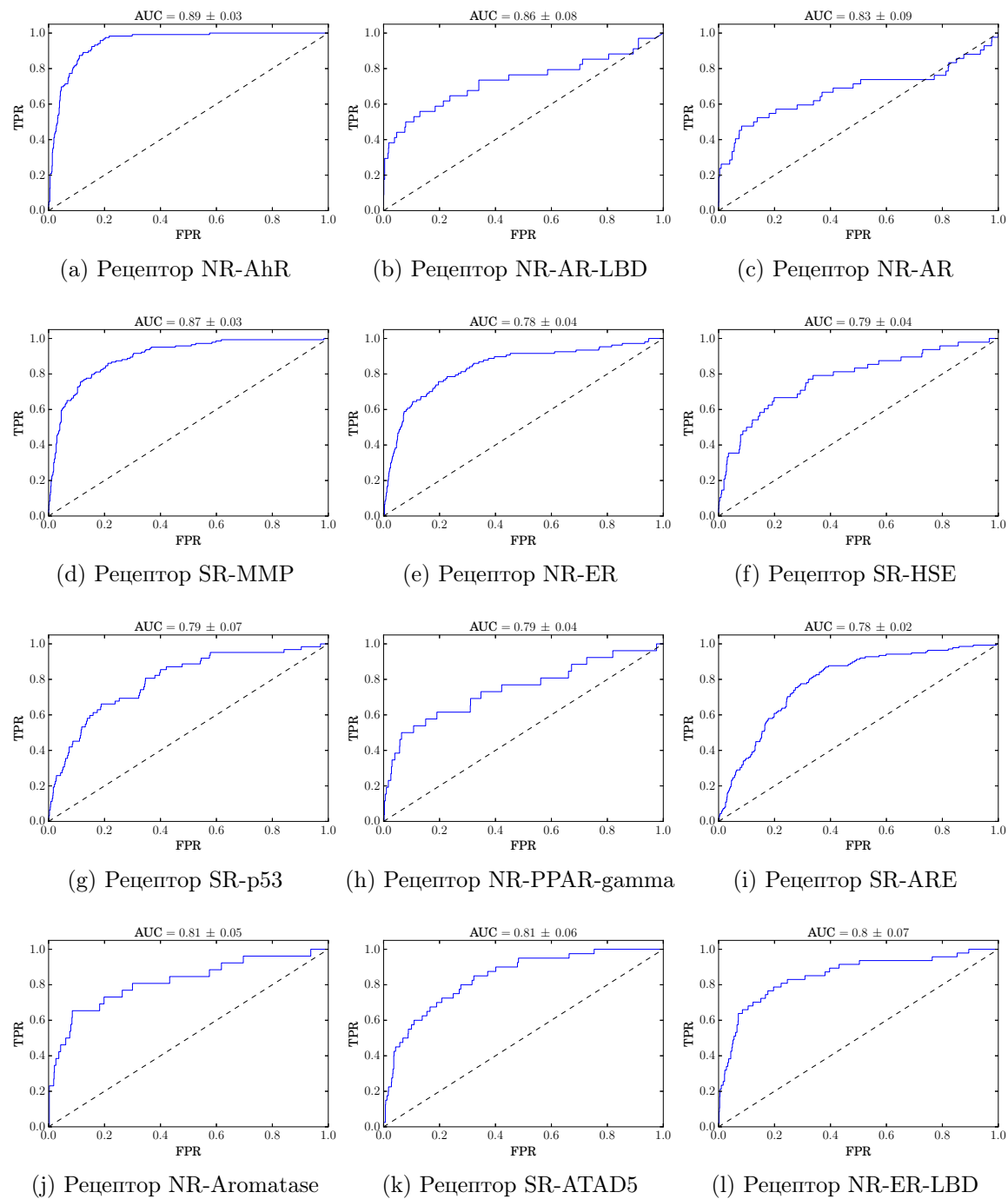


Рис. 4. ROC-кривая и значения функционала AUC для классов 1-12, метод Binary Relevance

На графиках (4) показаны ROC-кривые классов для одного из разбиений, построенные по предсказаниям Binary Relevance, а также значение функционала AUC. В таблице 3 приведено сравнение метода Binary Relevance с результатами из [17], для получения которых использовались те же данные, что и в данной работе. Сравнение результатов показывает, что логистическая регрессия алгоритм уступает в качестве классификации методу Random Forest. Для некоторых рецепторов эта разница значительна.

Для определения эффективности методов вычисляются значения метрик Hamming Loss, Subset Loss, Precision, Recall для каждого из разбиений $\mathfrak{D} = \mathfrak{L} \sqcup \mathfrak{T}$ на тестовую и контрольную выборку. Вычисляются средние значения и стандартные отклонения. Разбиения выполнены по методу Shuffle Split с размером тестовой выборки 0.1 и количеством разбиений 5 из-за разреженности данных. Используются функции потерь для PCC, аналогичные таковым для модельных данных. В эксперименте использованы только данные по рецепторам NR-AhR, NR-AR-LBD, NR-Aromatase. Используются только объекты со всеми тремя известными ответами.

Результаты сравнения PCC и BR представлены в таблице 4. Как и для модельных данных, заметно существенное улучшение Subset Loss для PCC (S). Также имеется незначительное улучшение Hamming Loss (H) для класса 2 (NR-AR-LBD).

Таблица 4. Сравнение алгоритмов на реальных данных. Рецепторы 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
H	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
P 1	0.79 ± 0.1	0.79 ± 0.1	0.79 ± 0.1	0.82 ± 0.1
P 2	0.91 ± 0.1	0.88 ± 0.1	0.91 ± 0.1	0.88 ± 0.1
P 3	0.76 ± 0.07	0.82 ± 0.09	0.78 ± 0.09	0.82 ± 0.08
R 1	0.17 ± 0.06	0.17 ± 0.06	0.15 ± 0.04	0.18 ± 0.05
R 2	0.22 ± 0.1	0.23 ± 0.1	0.24 ± 0.1	0.23 ± 0.1
R 3	0.1 ± 0.02	0.085 ± 0.02	0.071 ± 0.02	0.086 ± 0.02
S	0.32 ± 0.02	0.34 ± 0.02	0.46 ± 0.03	0.3 ± 0.03

Заключение

В работе применен алгоритм Probabilistic Classifier Chains для решения задачи предсказания взаимодействия рецепторов и лигандов. Алгоритм сравнивается с базовым алгоритмом, не учитывающим зависимости между классами. Вычислительный эксперимент показал, что как для модельных, так и для реальных данных PCC позволяет существенно улучшить показатели Subset Loss, т.е. качество предсказания всего вектора. При использовании Hamming Loss результаты сходны с результатами независимого классификатора. Предложена функция потерь для алгоритма PCC, позволяющая незначительно улучшить показатели Hamming Loss для отдельных классов.

Литература

- [1] R. DVORSKÝ V HORŇÁK and E. ŠTURDÍK. Receptor-ligand interaction and molecular modelling.
- [2] Tong Q Xie XQ Myint KZ, Wang L. Molecular fingerprint-based artificial neural networks qsar for ligand biological activity predictions. *Molecular Pharmaceutics*, 2012.
- [3] Xie XQ Myint KZ. Ligand biological activity predictions using fingerprint-based artificial neural networks (fann-qsar). *Methods Mol. Biol.*, 2015.
- [4] Bonnie Berger Vinay Pulim, Jadwiga Bienkowska. Lthreder: Prediction of extracellular ligand–receptor interactions in cytokines using localized threading. *Protein Science*, 2008.
- [5] Changhong Zhou Wenjun Zhang Zhengjun Cheng, Yuntao Zhang and Shibo Gao. Classification of 5-ht1a receptor ligands on the basis of their binding affinities by using pso-adaboost-svm.
- [6] Laurent Jacob and Jean-Philippe Vert. Protein–ligand interaction prediction: an improved chemogenomics approach. *BIOINFORMATICS*, 2008.
- [7] Peter Willett. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 1998.
- [8] Yusuke Komiyama et al. Masayuki Yarimizu, Cao Wei. Tyrosine kinase ligand-receptor pair prediction by using support vector machine. *Advances in Bioinformatics*, 2015.
- [9] Nagamani Sukumar Curt Breneman Scott Oloff†, Shuxing Zhang and Alexander Tropsha. Chemometric analysis of ligand receptor complementarity: Identifying complementary ligands based on receptor information (colibri). *J. Chem. Inf. Model.*, 2006.
- [10] M. Popova. Feature selection and multi-task prediction of biological activity for nuclear receptors. 11(1):111–112, 2015.
- [11] Jose Barranqueroa José Ramón Quevedoa Juan José del Coza Eyke Hüllermeierb Elena Montañesa, Robin Sengeb. Dependent binary relevance models for multi-label classification. *Pattern Recognition*, 2013.
- [12] Ivor W. Tsang Weiwei Liu. On the optimality of classifier chain for multi-label classification.
- [13] Geoff Holmes Eibe Frank Jesse Read, Bernhard Pfahringer. Classifier chains for multi-label classification.
- [14] Eyke H.O Krzysztof Dembczynski, Weiwei Cheng. Bayes optimal multilabel classification via probabilistic classifier chains. 2010.
- [15] Haytham Elghazel Maxime Gasse, Alex Aussem. On the optimality of multi-label classification under subset zero-one loss for distributions satisfying the composition property. 2015.
- [16] Eduardo F. Morales Pablo Hernandez-Leal Julio H. Zaragoza Pedro Larrañaga L. Enrique Sucar, Concha Bielza. Multi-label classification with bayesian network-based chain classifiers.
- [17] 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.