

# 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. 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.

## 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.

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] и [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 решена 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 активно-

---

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

сти 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:

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

Вводится предположение, что условное распределение  $P(\mathbf{X}|\mathbf{Y})$  принадлежит семейству экспоненциальных распределений.

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:

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

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  $\mathcal{L}$ :

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

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  $\mathbf{y}' \in \mathbf{Y}$ :

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

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 Precision и Recall для каждого класса  $j$  на контрольной выборке  $\mathcal{T}$  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  $P(\mathbf{y}|\mathbf{x})$ .

*Описать, как решается проблема с пропусками при обучении и придумать нормальный алгоритм (классификация уже имеющимися звеньями)*

*Описать, как производить разбиение*

*Описать, как выбирать порог логистической регрессии  $\frac{\lambda_+}{\lambda_-}$  по точке на ROC-кривой.*

## Часть 2. Работа с пропусками

Приведенный выше алгоритм PCC 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:  $\mathbf{x}' \leftarrow \|\mathbf{x}^T y_1 \dots y_{i-1}\|^{\square T}$

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 простого алгоритма для дальнейшего сравнения его с предлагаемым в статье. 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.

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 1 указано точное распределение ответов по классам.

**Таблица 1.** Количество связывающихся с рецепторами лигандов

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

Для определения эффективности данного метода вычисляются значения функционала AUC для каждого из разбиений  $\mathfrak{D} = \mathfrak{L} \sqcup \mathfrak{T}$  на тестовую и контрольную выборку. Разбиения выполнены по методу k-fold, где  $k = 5$ . Вычисляется среднее значение AUC, а также стандартное отклонение.

На графиках (1, 2) показаны ROC-кривые классов для одного из разбиений, а также значение функционала AUC.

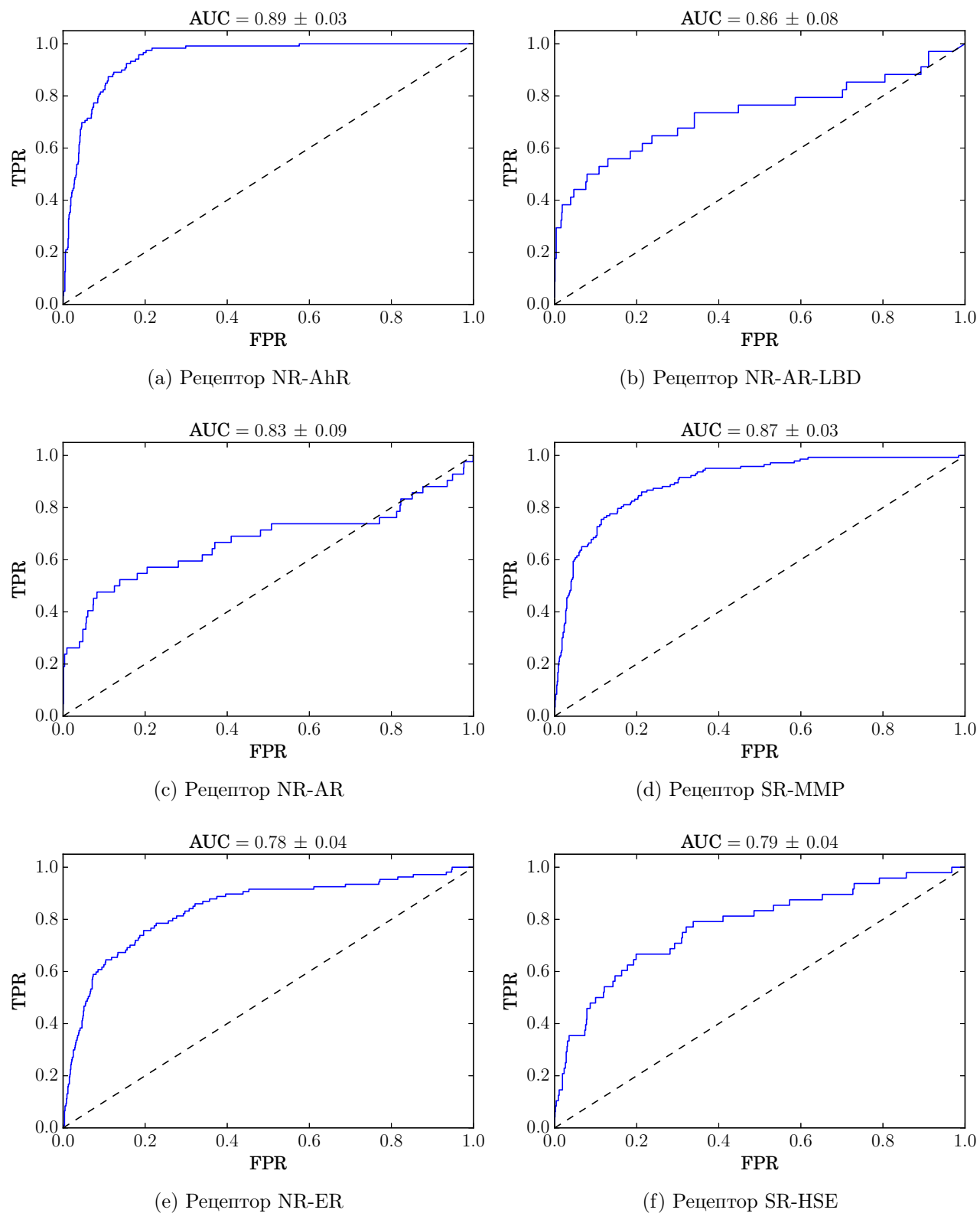
В таблице 2 приведено сравнение метода Binary Relevance с результатами из [17], для получения которых использовались те же данные и способ разбиения, что и в данной работе.

**Таблица 2.** Значение AUC для различных рецепторов и моделей классификации

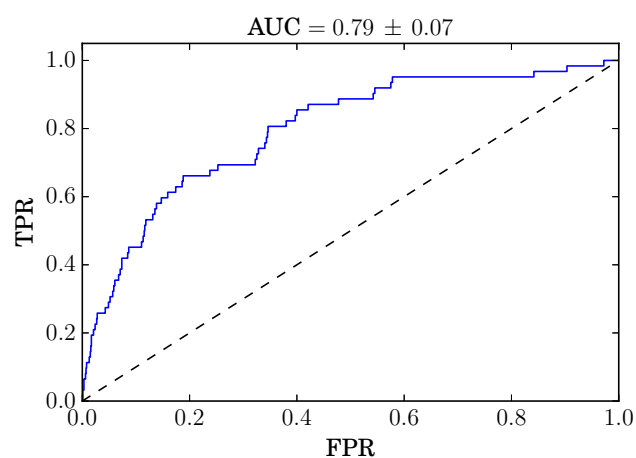
Рецептор	Binary Relevance	Random Forest [17]
NR-AhR	<b>0.83</b> $\pm$ 0.03	<b>0.93</b>
NR-AR-LBD	<b>0.86</b> $\pm$ 0.08	<b>0.88</b>
NR-AR	<b>0.83</b> $\pm$ 0.09	<b>0.83</b>
SR-MMP	<b>0.87</b> $\pm$ 0.03	<b>0.95</b>
NR-ER	<b>0.78</b> $\pm$ 0.04	<b>0.81</b>
SR-HSE	<b>0.79</b> $\pm$ 0.04	<b>0.86</b>
SR-p53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

Сравнение результатов показывает, что простой алгоритм уступает в качестве классификации методу Random Forest. Для некоторых рецепторов эта разница значительна.

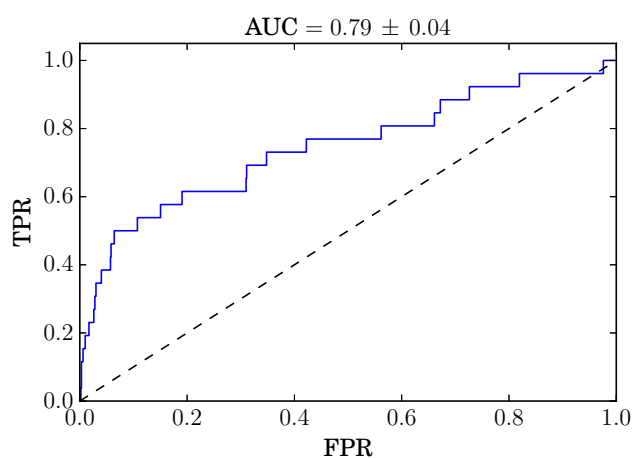
## Заключение



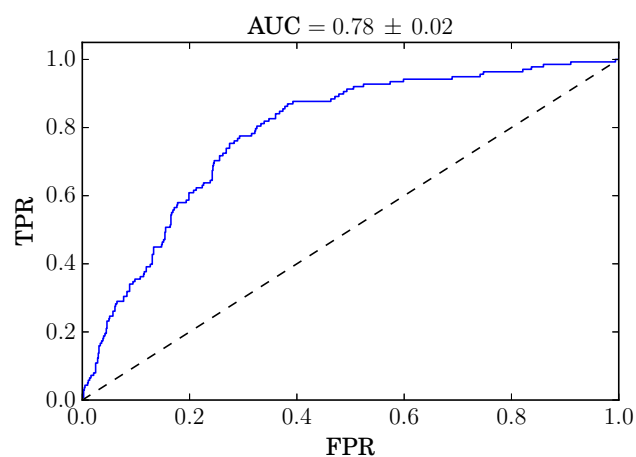
**Рис. 1.** ROC-кривая и значения функционала AUC для классов 1-6, метод Binary Relevance



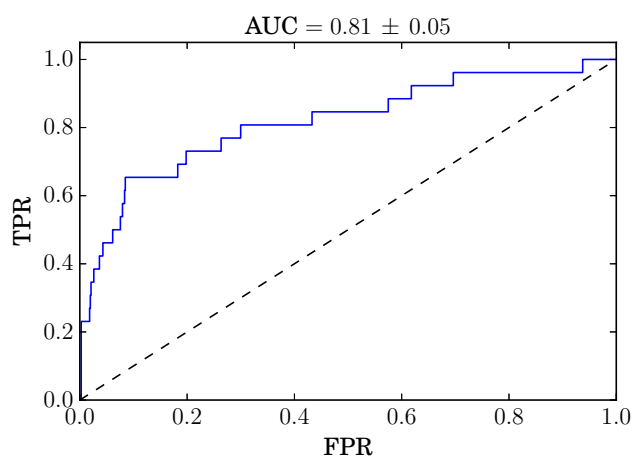
(a) Рецептор SR-p53



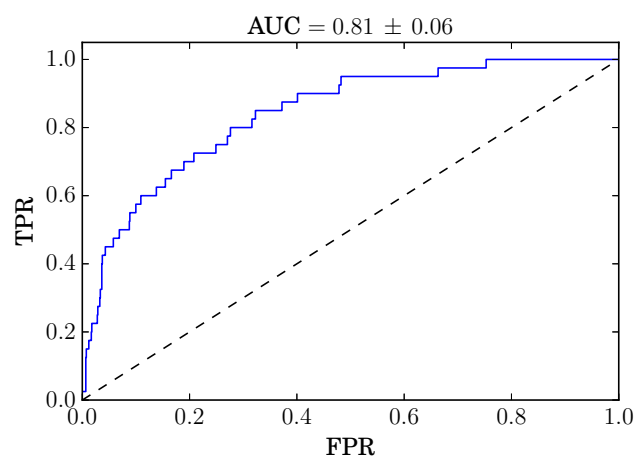
(b) Рецептор NR-PPAR-gamma



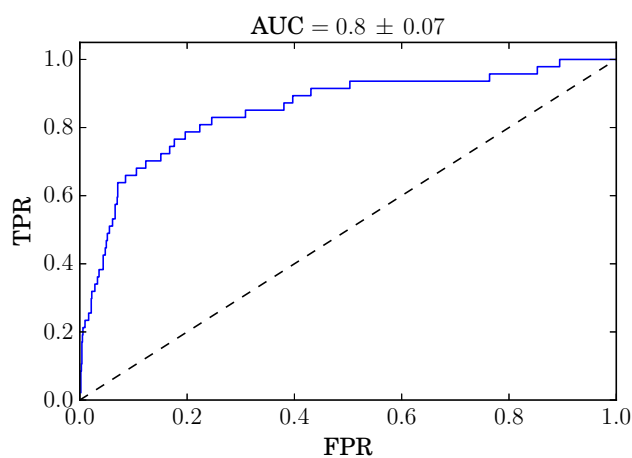
(c) Рецептор SR-ARE



(d) Рецептор NR-Aromatase



(e) Рецептор SR-ATAD5



(f) Рецептор NR-ER-LBD

**Рис. 2.** ROC-кривая и значения функционала AUC для классов 6-12, метод Binary Relevance



## 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.