Бэггинг нейронных сетей в задаче анализа биологической активности ядерных рецепторов*

М. Р. Владимирова, М. С. Попова
mrvladimirova@gmail.com; popova@gmail.com
Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., д. 9

Работа посвящена решению проблемы повышения качества многозадачной классификации с помощью нейросетевой модели. Улучшение модели решения задачи проводится многозадачной моделью двухслойной нейронной сети. Рассматриваются две функции потерь: квадратичная и кросс-энтропийная. Для получения более точного результата в работе рассматривается композиция базовых классификаторов — бэггинг нейронных сетей. Сравнение моделей проводится с помощью вычислительного эксперимента на реальных данных, описывающих взаимодействия рецепторов и лиганд.

Ключевые слова: *клеточные рецепторы; биологическая активность; двухслойная нейронная сеть; бэггинг; многозадачность; разработка лекарств; кросс-энтропийная функция*

DOI: 10.21469/22233792.2.3.06

1 Введение

Рассматривается проблема многозадачной классификации на данных, описывающих взаимодействие ядерных рецепторов. Ядерные рецепторы представляют собой класс находящихся в клетках белков. Рецепторы влияют на транскрипцию генов: регулируют развитие, гомеостаз и обмен веществ в организме. Регулирование происходит в основном тогда, когда рецептор и лиганд — молекула, воздействующая на поведение рецептора, — взаимодействуют. Требуется предсказать, будет ли объект относиться к определенному классу, т. е. будет ли взаимодействовать данный лиганд с определенным рецептором. Проблема построения адекватных математических моделей для предсказания лиганд-рецепторного взаимодействия на основании данных о структурах химических соединений является актуальной задачей в фармакологии [1–4]. С помощью моделей проводится предварительная оценка характера взаимодействия лиганд и рецепторов, что позволяет снизить количество лабораторных экспериментов, необходимых для выявления активных лиганд.

Существуют два подхода к решению данной проблемы. Один из подходов заключается в компьютерном моделировании взаимодействия молекул, основанном на законах молекулярной динамики [5]. Такой способ является трудоемким и неприменим в случаях, когда точная трехмерная структура рецептора или лиганда неизвестна [6]. Второй подход — использование методов, основанных на статистике и машинном обучении. В литературе такой подход получил общее название «поиск количественных соотношений структура–свойство», или «Quantitative Structure–Activity Relationship» [7]. Модели, связывающие структуру лиганд с их биологической активностью, показали свою способность к предсказыванию лиганд-рецепторного взаимодействия [8, 9]. Точность модели машинного обучения зависит от размера обучающей выборки, поэтому для построения точных моделей необходим достаточный объем выборки. Несмотря на то что для некоторых рецепторов

*Проект поддержан грантом РФФИ № 16-07-01155.

уже проведено немало лабораторных экспериментов, данных о многих рецепторах оказывается недостаточно [10, 11]. Однако экспертные знания в области биохимии и фармакологии дают основания полагать, что факты связывания одних и тех же молекул с разными рецепторами не являются независимыми. Это означает, что можно компенсировать недостаток известных лиганд для данной цели наличием известных лиганд для подобных целей, используя многозадачное предсказание.

В данной работе решается набор взаимосвязанных или схожих задач обучения одновременно, с помощью алгоритмов обучения, имеющих схожее внутреннее представление, т. е. решается проблема многозадачной классификации. Информация о сходстве задач между собой позволяет совершенствовать алгоритм обучения и повышать качество решения основной задачи. Моделью классификации, позволяющей строить предсказания для группы рецепторов, предлагается использовать двухслойную нейронную сеть. Искусственные нейронные сети — эффективный инструмент решения исследовательских задач [8, 12–14]. Нейронные сети обладают уникальными особенностями, которые делают их надежными для решения задач с многомерными входными данными. Например, сети устойчивы к изменениям во входных данных [15], являются мультитасковыми, т. е. могут одновременно решать несколько задач [16], обучаются на всей выборке, не фрагментируя ее [17, 18].

Для повышения качества предсказаний лиганд-рецепторных взаимодействий предлагается использовать композицию двухслойных нейронных сетей. Одним из способов получения композиции классификаторов является использование бэггинга (bootstrap aggregating) [19]. Бэггинг генерирует из элементов обучающей выборки размера n семейство подвыборок размера n с помощью процедуры бутстрэп (bootstrap). Процедура основана на выборках с возвращениями, т. е. некоторые объекты могут встречаться в подвыборке более одного раза, а другие — отсутствовать. На каждой подвыборке настраивается классификатор. Ответы классификаторов агрегируются путем простого голосования. Бэггинг над базовыми алгоритмами позволяет увеличить точность и повысить устойчивость модели [20].

При решении задачи многоклассовой классификации на выходе нейронной сети необходимо получить вероятность принадлежности объекта каждому из классов. Рассмотрены две дифференцируемые функции потерь: квадратичная и кросс-энтропийная. Первая — сумма квадратов разности между истинным и восстановленным значениями. Чтобы знать суммарное число несовпадений между восстановленными метками классов и фактическими, используется кросс-энтропийная функция потерь — функция наибольшего правдоподобия в задаче логистической регрессии.

В работе был проведен вычислительный эксперимент на реальных данных, в ходе которого базовый алгоритм, двухслойная нейронная сеть, сравнивался с бэггингом над базовыми алгоритмами. Сравнение проводилось по значению функционала AUC (area under curve).

2 Постановка задачи

В задаче исследуется взаимодействие N лиганд с M рецепторами. Дана выборка $\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, \ldots, N\}$, состоящая из N пар объект–ответ.

Объектами $\mathbf{x}_i \in \mathbb{R}^K$ являются вектора признаковых описаний, в которых хранятся числовые свойства лиганд. Значения компонент вектора ответа $\mathbf{y}_i \in \{0, 1\}^M$ показывают, есть ли связь лиганда, соответствующего описанию $\mathbf{x}_i$, с различными рецепторами. Если реальный эксперимент не проводился или не дал адекватных результатов, то в ответе стоит пропуск. Назовем рецептор, взаимодействие с которым описывается m-м элементом

вектора ответа $\{y_i^m\}_{i=1}^N \in \{0,1\}$, m-рецептором, где $m \in \{1,\ldots,M\}$. Если лиганд с описанием $\mathbf{x}_i$ активирует m-рецептор, то $y_i^m = 1$, если не активирует — $y_i^m = 0$. Предположим, что $\mathbf{y}_i$ является реализацией случайного вектора, каждая компонента которого имеет распределение Бернулли. Исследуем взаимодействие каждого рецептора в разных задачах бинарной классификации. Пусть m-рецептору соответствует m-я задача, тогда решим одновременно M задач бинарной классификации, построив единую модель.

Базовые алгоритмы выбираются из класса двухслойных нейронных сетей:

$$\mathbf{z}(\mathbf{x},\boldsymbol{\theta}) = \mathbf{W}_2^{\mathsf{T}}\tanh(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}) : \mathbb{R}^K \to \mathbb{R}^H; \tag{1}$$

$$\mathbf{f}(\mathbf{x},\boldsymbol{\theta}) = \frac{1}{1+\mathbf{exp}\big(-\mathbf{z}(\mathbf{x},\boldsymbol{\theta})\big)} : \mathbb{R}^K \to [0,1]^M, \tag{2}$$

где $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W}_1^{\mathsf{T}}|\mathbf{W}_2^{\mathsf{T}})$ — вектор параметров двухслойной сети.

Значения признаков объекта $\mathbf{x}$ поступают на вход первому входному слою сети с весовой матрицей $\mathbf{W}_1$. Выходы первого слоя поступают на вход второму с весовой матрицей $\mathbf{W}_2$ — скрытому слою. Ответы на выходном слое интерпретируются как оценки вероятности того, что лиганд $\mathbf{x}$ связывается с рецепторами соответствующих задач:

$$\mathbf{f}(\mathbf{x},\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{P}(y_1 = 1|\mathbf{x},\boldsymbol{\theta}) \\ \mathsf{P}(y_2 = 1|\mathbf{x},\boldsymbol{\theta}) \\ \vdots \\ \mathsf{P}(y_M = 1|\mathbf{x},\boldsymbol{\theta}) \end{bmatrix}. \tag{3}$$

Выборка $\mathfrak{D}$ разделяется на две подвыборки — обучающую и контрольную. Для формирования бутстрэп-выборок $\mathfrak{L}_\ell$, $\ell = \{1,\ldots,L\}$, из обучающей выборки $\mathfrak{L}$ случайным образом отбирается несколько подмножеств, содержащих такое же количество элементов, как и исходное. Поскольку отбор производится случайно, набор элементов в этих выборках будет различным: некоторые из них могут быть отобраны по несколько раз, а другие — ни разу. Доля уникальных элементов в полученных выборках в среднем равна 0,56. На каждой из L выборок обучается базовый классификатор. Ответы классификаторов агрегируются путем простого голосования.

Моделью классификации $\mathbf{a}$, решающую одновременно M задач, назовем композицию базовых алгоритмов

$$\mathbf{a}(\mathbf{x},\boldsymbol{\theta},L) = \sum_{\ell=1}^{L} \pi_\ell \mathbf{f}(\mathbf{x},\boldsymbol{\theta}^\ell), \tag{4}$$

где $\pi_\ell = 1/L$ — веса базовых классификаторов; $\mathbf{f}(\mathbf{x},\boldsymbol{\theta}^\ell)$ — базовый алгоритм; $\boldsymbol{\theta}^\ell$ — вектор параметров базового алгоритма, вычисленного на подвыборке $\mathfrak{L}_\ell$.

Рассмотрим две задачи: линейную и логистическую регрессию. Каждой задаче соответствует функция ошибки, $\mathcal{L}_1$ и $\mathcal{L}_2$. Определим для каждой суммарную функцию потерь Q на некоторой подвыборке $\mathfrak{U}$ исходной выборки $\mathfrak{D}$ следующим образом:

$$\mathcal{L}_1(\boldsymbol{\theta},\mathbf{x}_i,\mathbf{y}_i) = \frac{1}{2}\sum_{m=1}^{M}\big(a^m(\mathbf{x}_i,\boldsymbol{\theta},L) - y_i^m\big)^2, \tag{5}$$

где $a^m(\mathbf{x}_i,\boldsymbol{\theta},L)$ — ответ классификатора на объекте $\mathbf{x}_i$ в m-й задаче, m-я компонента вектора $\mathbf{a}(\mathbf{x}_i,\boldsymbol{\theta},L)$ (4):

$$\mathcal{L}_2(\boldsymbol{\theta},\mathbf{x}_i,\mathbf{y}_i) = -\sum_{m=1}^{M} y_i^m \log \mathsf{P}(y_i^m = 1|\mathbf{x}_i,\boldsymbol{\theta}) + (1-y_i^m)\log \mathsf{P}(1-y_i^m = 1|\mathbf{x}_i,\boldsymbol{\theta}); \tag{6}$$

$$Q(\boldsymbol{\theta}, L|\mathfrak{U}) = \sum_{i=1}^{|\mathfrak{U}|} \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}_i, \mathbf{y}_i),\ \ \mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2\}.$$

Для нахождения оптимальных параметров $\hat{\mathbf{w}}$ и $\hat{L}$ модели $\mathbf{a}$ требуется решить задачу минимизации функции ошибки на обучающей выборке:

$$\hat{\boldsymbol{\theta}}, \hat{L} = \operatorname*{argmin}_{\boldsymbol{\theta}, L} Q(\boldsymbol{\theta}, L|\mathfrak{L}). \tag{7}$$

Для дополнительной оценки качества классификации будем вычислять значения функционала AUC на контрольной выборке для каждого класса по принципу один против всех и визуализировать полученные результаты с помощью ROC (receiver operating characteristic) кривых.

3 Оптимизация модели

Проанализируем построенную модель (1), (2), (4) с помощью декомпозиции ошибки Q на компоненты смещения и разброса (bias-variance decomposition) [21, 22]. Рассмотрим без потери общности декомпозицию функции ошибки для одной компоненты объекта выборки и одной задачи $a^m(x) = a(x)$ (4).

3.1 Квадратичная функция потерь

Пусть x — объект; y — истинная зависимость от объекта x; $f(x)$ — некоторый алгоритм, аппроксимирующий y. Квадратичной функции потерь (5) соответствует квадратичный риск

$$R(f) = \mathsf{E}_{x,y}\Big[\big(y - f(x)\big)^2\Big]. \tag{8}$$

Минимум среднеквадратичного риска достигается на функции, возвращающей условное матожидание ответа на фиксированном объекте. В случае бинарной классификации условие на минимум записывается следующим образом:

$$f^*(x) = \mathsf{E}[y|x] = \mathsf{P}(y = 1|x) = \operatorname*{argmin}_f R(f).$$

В работе рассматривается вероятностная модель (3). Вероятностная регрессионная модель лучше описывает предсказание вероятности биномиально распределенных величин в смысле среднеквадратичной ошибки, чем модель бинарной классификации:

$$\begin{aligned}\mathsf{E}\Big[\big(y - f(x)\big)^2\big|x\Big] &= \mathsf{E}\Big[\Big(\big(y - \mathsf{E}[y|x]\big) + \big(\mathsf{E}[y|x] - f(x)\big)\Big)^2\Big] = \\ &= \mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)^2\big|x\Big] + \Big(\mathsf{E}[y|x] - f(x)\Big)^2 + 2\,\mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)\big|x\Big] \cdot \Big(\mathsf{E}[y|x] - f(x)\Big) = \\ &= \mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)^2\big|x\Big] + \Big(\mathsf{E}[y|x] - f(x)\Big)^2 \geqslant \mathsf{E}\Big[\big(y - \mathsf{E}[y|x]\big)^2\big|x\Big].\end{aligned}$$

Опишем зависимость среднеквадратичного риска (8) от выборки $\mathfrak{L}$ для композиции алгоритмов (4). Основной мерой качества алгоритма $a(x)$ возьмем усредненный по всем выборкам среднеквадратичный риск:

$$\mathcal{L}(a) = \mathsf{E}_{\mathfrak{L}}\Big[\mathsf{E}_{x,y}\Big[\big(y - a(x, \mathfrak{L})\big)^2\Big]\Big].$$

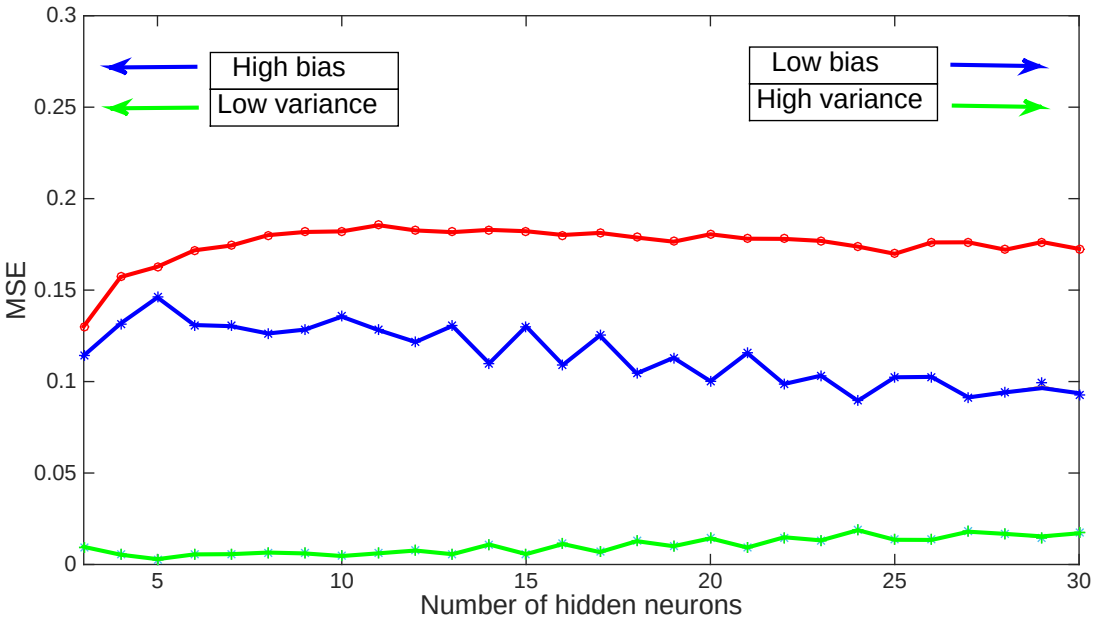

Рис. 1 Смещение и дисперсия базового алгоритма в зависимости от количества нейронов на скрытом слое

Для квадратичной функции ошибки для любого a $\mathcal{L}(a)$ представима в виде суммы из трех слагаемых [23]:

$$\mathcal{L}(a) = \mathsf{E}_{x,y}\Big[\big(y - \mathsf{E}\,[y|x]\big)^2\Big] + \mathsf{E}_{x,y}\Big[\Big(\mathsf{E}_{\mathfrak{L}}\big[a(x,\mathfrak{L})\big] - \mathsf{E}[y|x]\Big)^2\Big] + \\ + \mathsf{E}_{x,y}\Big[\mathsf{E}_{\mathfrak{L}}\Big[\Big(a(x,\mathfrak{L}) - \mathsf{E}_{\mathfrak{L}}\big[a(x,\mathfrak{L})\big]\Big)^2\Big]\Big]. \quad (9)$$

Первая компонента равна ошибке идеального алгоритма и описывает шум в данных. Невозможно построить алгоритм, имеющий меньшее ожидание ошибки. Вторая компонента характеризует смещение (bias) метода обучения, т. е. отклонение среднего ответа обученного алгоритма от ответа идеального алгоритма. Третья компонента характеризует дисперсию (variance), т. е. разброс ответов обученных алгоритмов относительно среднего ответа.

На рис. 1 показана визуализация зависимости смещения (синей линией) и дисперсии (зеленой линией) базового алгоритма от размерности пространства параметров, количества нейронов на скрытом слое. Также красной линией показана суммарная ошибка в зависимости от количества нейронов на скрытом слое. С увеличением количества скрытых нейронов смещение уменьшается, а дисперсия увеличивается.

Теорема 1. *Смещение композиции, полученной с помощью бэггинга, совпадает со смещением одного базового алгоритма* (2).

Доказательство.

$$\mathsf{E}_{x,y}\left[\left(\mathsf{E}_{\mathfrak{L}}\left[\frac{1}{L}\sum_{\ell=1}^{L} f(x,\mathfrak{L}_\ell)\right]-\mathsf{E}[y|x]\right)^2\right]=\mathsf{E}_{x,y}\left[\left(\frac{1}{L}\sum_{\ell=1}^{L}\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)-\mathsf{E}[y|x]\right]\right)^2\right]=$$
$$=\mathsf{E}_{x,y}\left[\left(\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)-\mathsf{E}[y|x]\right]\right)^2\right]=\mathsf{E}_{x,y}\left[\left(\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)\right]-\mathsf{E}[y|x]\right)^2\right]. \quad (10)$$

Таким образом, бэггинг не ухудшает смещенность модели. ■

Теорема 2. *Дисперсия композиции в L раз меньше дисперсии отдельных алгоритмов.*

Доказательство. Дисперсия композиции, построенной с помощью бэггинга, состоит из дисперсии одного базового алгоритма и корреляции между базовыми алгоритмами:

$$\mathsf{E}_{x,y}\left[\mathsf{E}_{\mathfrak{L}}\left[\left(\frac{1}{L}\sum_{\ell=1}^{L} f(x,\mathfrak{L}_\ell)-\mathsf{E}_{\mathfrak{L}}\left[\frac{1}{L}\sum_{\ell=1}^{L} f(x,\mathfrak{L}_\ell)\right]\right)^2\right]\right]=$$
$$=\frac{1}{L}\,\mathsf{E}_{x,y}\left[\mathsf{E}_{\mathfrak{L}}\left[\left(f(x,\mathfrak{L}_\ell)-\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)\right]\right)^2\right]\right]+$$
$$+\frac{L-1}{L}\,\mathsf{E}_{x,y}\left[\mathsf{E}_{\mathfrak{L}}\left[\left(f(x,\mathfrak{L}_\ell)-\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_\ell)\right]\right)\left(f(x,\mathfrak{L}_k)-\mathsf{E}_{\mathfrak{L}}\left[f(x,\mathfrak{L}_k)\right]\right)\right]\right]. \quad (11)$$

Если базовые алгоритмы некоррелированы, то дисперсия композиции в L раз меньше дисперсии отдельных алгоритмов. Поскольку нейронные сети относятся к неустойчивым моделям, корреляция алгоритмов отсутствует. ■

Таким образом, из теорем 1 и 2 следует, что бэггинг обеспечивает повышение точности.

3.2 Кросс-энтропийная функция потерь

Для случайных величин, имеющих распределение Бернулли, задается кросс-энтропийная функция потерь (6). Выразим данную функцию потерь через расстояние Кульбака–Лейблера. Рассмотрим задачу бинарной классификации $y=\{0,1\}$. Пусть p — истинная вероятность $\mathsf{P}(y=1|x)$ принадлежности объекта x к классу $y=1$; f — гипотетическая вероятность $\mathsf{P}(y=1|x)$, полученная с помощью алгоритма, аппроксимирующего y (3). Тогда расстояние Кульбака–Лейблера выражается следующим образом:

$$D_{\mathrm{KL}}(p,f)=p\ln\frac{p}{f}+(1-p)\ln\frac{1-p}{1-f}.$$

Обозначим $f^*(x)$ решение задачи

$$f^*(x)=\operatorname*{argmin}_{f\in[0,1]}\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\big(y,f\big)\right]. \quad (12)$$

Тогда получаем среднее геометрическое:

$$\ln\frac{f^*(x)}{1-f^*(x)}=\mathsf{E}_{x,y}\left[\ln\frac{f(x)}{1-f(x)}\right],$$

откуда

$$f^*(x)=\frac{1}{Z}\exp\left(\mathsf{E}_{x,y}\left[\ln f(x)\right]\right),$$

где Z — нормировочная константа, не зависящая от y.

Основной мерой качества алгоритма $a(x)$ возьмем усредненное по всем выборкам расстояние Кульбака–Лейблера:

$$\mathcal{L}(a) = \mathsf{E}_{\mathfrak{L}}\left[\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\left(y, a(x, \mathfrak{L})\right)\right]\right]. \tag{13}$$

Для решения задачи (12) необходимо разложить (13) на шум, смещение и дисперсию, как это было сделано для квадратичной функции потерь (9).

Теорема 3. *Ошибкой идеального алгоритма является энтропия от истинной вероятности $H(p)$.*

Доказательство. Расстояние между истинными ответами y и истинной вероятностью принадлежности объекта к классам p будет являться шумом в данных (ошибкой идеального алгоритма):

$$\begin{aligned}\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}(y,p)\right] &= \mathsf{E}_{x,y}\left[y\ln\frac{y}{p} + (1-y)\ln\frac{1-y}{1-p}\right] = \\ &= \mathsf{E}_{x,y}\left[y\ln y - y\ln p + (1-y)\ln(1-y) - (1-y)\ln(1-p)\right] = \\ &= -p\ln p - (1-p)\ln(1-p) = H(p),\end{aligned}$$

где $H(p)$ — функция энтропии. ■

Утверждение 1. *Смещение B и дисперсия V для функции ошибки $\mathcal{L}$ выражаются следующим образом* [22]*:*

$$\begin{aligned} B &= \mathcal{L}\left(p, a^*(x)\right), & a^*(x) &= \operatorname*{argmin}_{a\in[0,1]} \mathcal{L}(y,a); \\ V &= \mathsf{E}_{x,y}\left[\mathcal{L}\left(a^+(x), a^*(x)\right)\right], & a^+(x) &= \operatorname*{argmin}_{a\in[0,1]} \mathcal{L}(p,a). \end{aligned}$$

Из теоремы 3 и утверждения 1 получаем, что выражение (13) представляется в виде суммы трех слагаемых:

$$\mathcal{L}(a) = \mathsf{E}_{\mathfrak{L}}\left[\mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\left(y, a(x,\mathfrak{L})\right)\right]\right] = H(p) + D_{\mathrm{KL}}\left(p, a^*(x)\right) + \mathsf{E}_{x,y}\left[D_{\mathrm{KL}}\left(\mathsf{E}_{\mathfrak{L}}\left[a(x,\mathfrak{L})\right], a^*(x)\right)\right].$$

Тогда смещение бэггинга совпадает со смещением одного базового алгоритма, а дисперсия бэггинга уменьшается [22].

Таким образом, проводя сравнение разложений ошибки между композицией алгоритмов и одним базовым алгоритмом, получили, что для обеих функций потерь $\mathcal{L}_1$ и $\mathcal{L}_2$ выполняется равенство смещений и уменьшение дисперсий. Это означает, что результаты бэггинга нейронных сетей должны быть точнее, чем отдельной нейронной сети. Подтвердим вышеизложенные теорические выкладки вычислительным экспериментом.

3.3 Нахождение параметров модели

Оптимизация вектора параметров $\boldsymbol{\theta}$, минимизирующего суммарную функцию потерь (7) по обучающей выборке $\mathfrak{L}$, проводится модифицированным методом обратного распространения ошибки. Псевдокод алгоритма в представлен в Алгоритме 1.

Алгоритм 1 Модифицированный метод обратного распространения ошибки

Вход: выборка $\mathfrak{D} = \left\{(\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N\right\}$, количество циклов C, число нейронов в скрытом слое H, темп обучения сети η, модель с заданными функциями активации на первом α_1 и втором слоях α_2;

Выход: весовые параметры w_{jh}, w_{hm};

инициализировать веса w_{jh}, w_{hm};
задать $k = 0$;
повторять
выбрать объект $\mathbf{x}_i$ из $\mathfrak{D}$;
прямой ход:
вычислить значение функции на скрытом слое $u_i^h := \alpha_{1h}\left(\sum_{j=0}^n w_{jh}x_i^j\right),\ h = 1, \ldots, H$,
вычислить значение функции на выходном слое $f_i^m := \alpha_{2m}\left(\sum_{h=0}^H w_{hm}u_i^h\right),\ m = 1, \ldots, M$,
если есть результаты экспериментов: $y_i^m = 0$ или $y_i^m = 1$ **то**
вычислить значение ошибки на выходном слое ε_i^m,
функция ошибки квадратичная: $\varepsilon_i^m := f_i^m - y_i^m$,
функция ошибки кросс-энтропийная: $\varepsilon_i^m := -y_i^m / f_i^m - (1 - y_i^m) / (1 - f_i^m)$;
иначе
обработка пропусков $\varepsilon_i^m := 0$,
обратный ход:
вычислить значение ошибки на скрытом слое $\varepsilon_i^h := \sum_{m=1}^M \varepsilon_i^m \alpha'_{2m} w_{hm},\ h = 1, \ldots, H$,
α'_2 — функция, обратная к функции активации;
градиентный шаг:
$w_{hm} := w_{hm} - \eta\varepsilon_i^m \alpha'_{2m} u_i^h,\ h = 0, \ldots H,\ m = 1, \ldots, M$,
$w_{jh} := w_{jh} - \eta\varepsilon_i^m \alpha'_{1h} x_i^j,\ j = 0, \ldots n,\ h = 1, \ldots, H$;
$k := k + 1$;
пока $k < C$;

4 Вычислительный эксперимент

Выборка $\mathfrak{D}$ состоит из описания взаимодействия $N = 8513$ лиганд с $M = 12$ рецепторами: NR-AhR, NR-AR-LBD, NR-AR, SR-MMP, NR-ER, SR-HSE, SR-p53, NR-PPAR-gamma, SR-ARE, NR-Aromatase, SR-ATAD5 и NR-ER-LBD. Каждый объект описан $K = 185$ признаками. Биологическая активность выражается бинарным значением ответов: 1 — есть взаимодействие; 0 — нет взаимодействия. Если реальный эксперимент не проводился или не дал результатов, то в ответе стоит пропуск. На рис. 2 указано распределение объектов по классам. Около половины объектов — с известым бинарным ответом. Доля полностью размеченных объектов составляет 16% исходной выборки. Объект с пропуском в ответе не участвовал в тестировании.

Проведен вычислительный эксперимент на реальных данных, представленных в выборке $\mathfrak{D}$. Цель эксперимента — проверить адекватность работы базового алгоритма; получить оценки качества, необходимые для сравнения с предложенной в работе моделью классификации; сравнить качество результатов, полученных с помощью базового алгоритма и предложенной модели классификации.

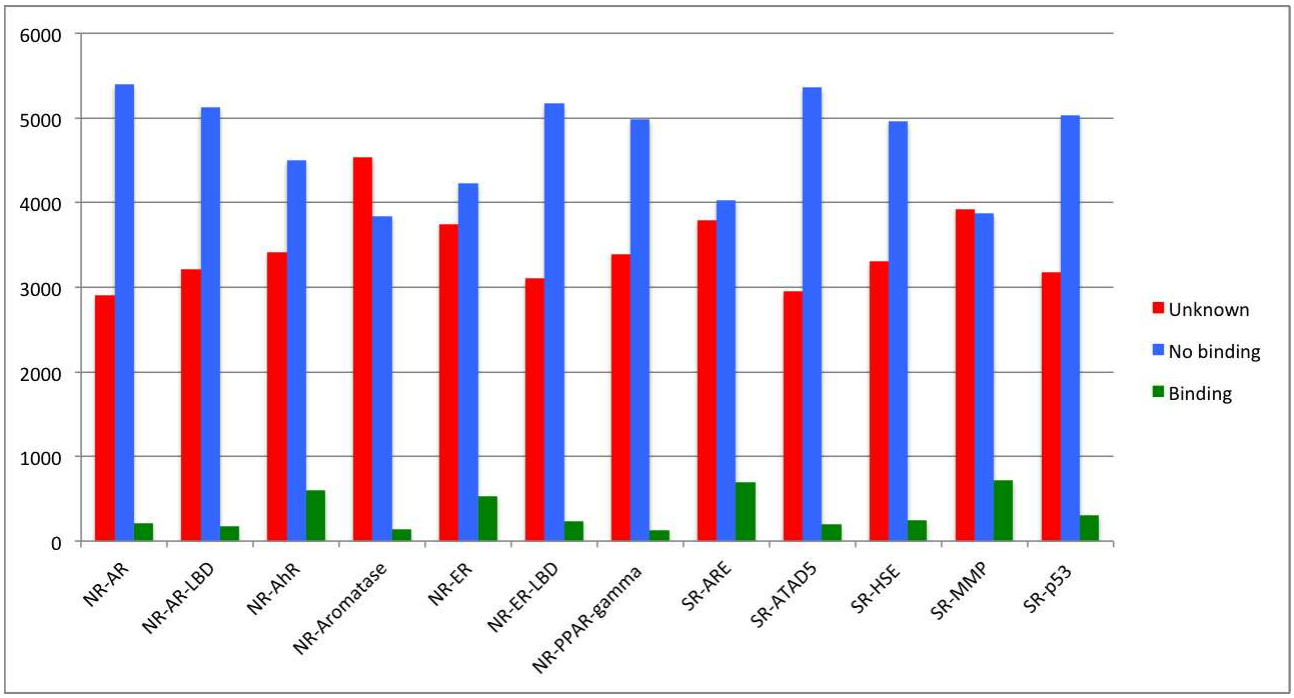

Рис. 2 Количество связывающихся лиганд для каждого рецептора

4.1 Базовый эксперимент

Для оценки качества результата была использована кросс-валидация обучающей выборки на 5 непересекающихся блоков. Для проверки качества алгоритма использованы ROC-кривые. На контрольных выборках вычислены значения функционала AUC. Для каждого рецептора синий линией на рис. 3 изображена ROC-кривая с вычисленным значением AUC.

Таким образом, нейронная сеть показала свою способность предсказывать биологическую активность лиганд и рецепторов.

4.2 Настройка параметров

Для улучшения качества классификации нейронной сети проведена настройка параметра H числа нейронов на скрытом слое. Значение функционала AUC вычислялось в зависимости от числа нейронов из промежутка от 1 до 100 с шагом, равным 5. На рис. 4 приведены зависимости для первых трех рецепторов. Результаты на большинстве рецепторов незначительно меняются в зависимости от H, но на некоторых точность классификации увеличивается, как для рецептора NR-AhR. Возьмем значение $H = 100$, при котором значение функционала AUC стабилизируется и остается примерно константой.

Параметр $\hat{L}$ находится с помощью анализа графика зависимости значения функционала AUC от количества разбиений в модели классификации (см. (7)). На графике для рецептора NR-AhR на рис. 5 видно, что с увеличением числа разбиений значение AUC растет, но с определенного момента значение остается константным. Проанализировав графики для всех 12 рецепторов, выбираем $L = 100$, при котором для каждого рецептора значение AUC на графике становится постоянным.

4.3 Бэггинг

Проведен вычислительный эксперимент для предложенной модели бэггинга с оптимальными параметрами. Результаты эксперимента показаны на рис. 3. Красной линией

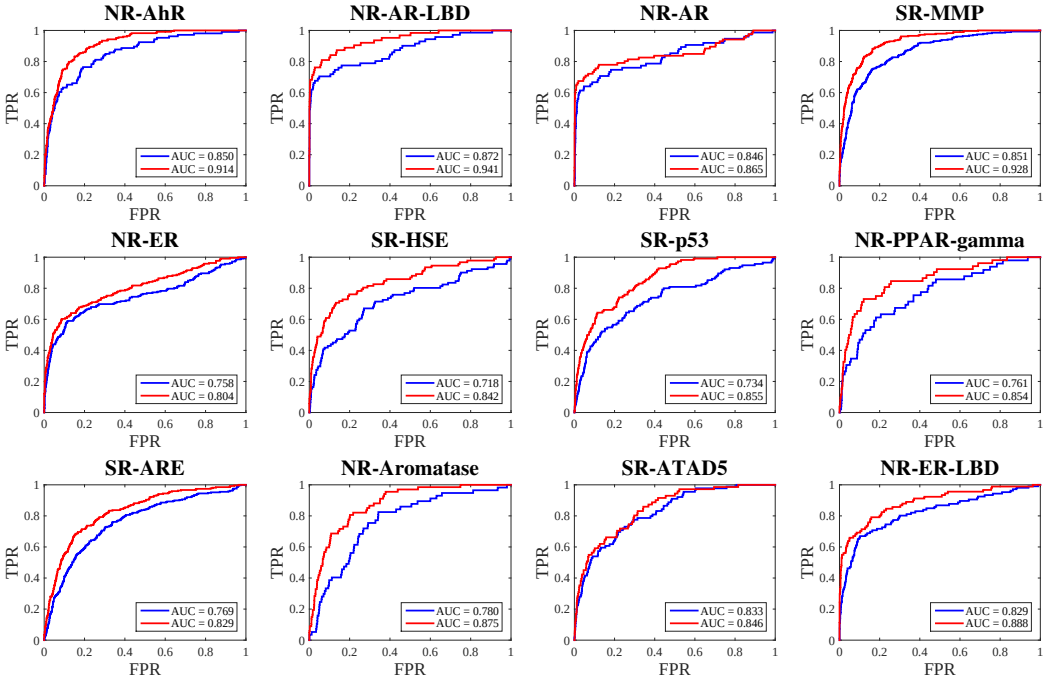

Рис. 3 ROC-кривые базового алгоритма и бэггинга

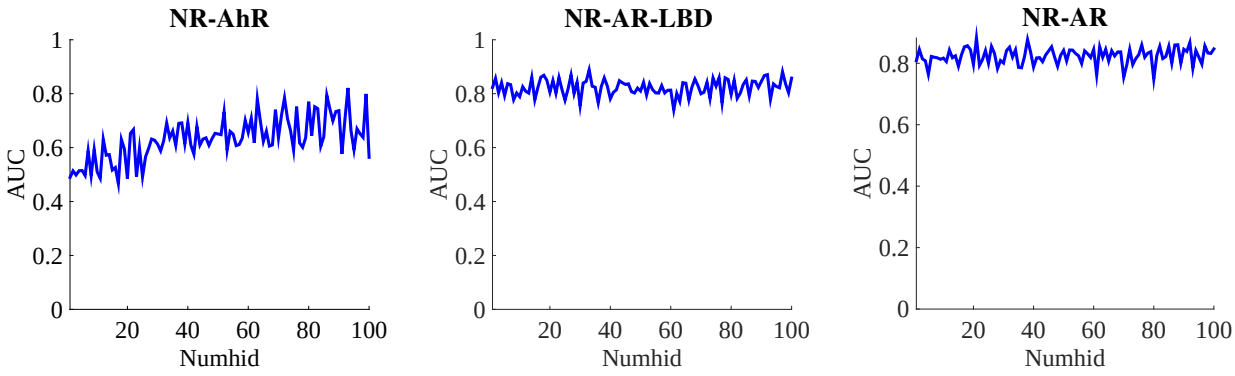

Рис. 4 Графики зависимости значения функционала AUC от количества нейронов на скрытом слое

изображены ROC-кривые бэггинга с вычисленным значением AUC. Площадь под кривой базового алгоритма для каждого рецептора меньше площади под соответствующей кривой бэггинга. Сравнение значений функционала AUC, полученного с помощью базового алгоритма нейронной сети и предложенного алгоритма бэггинга нейронных сетей, приведено в табл. 1. Сравнение проведено также между двумя функциями потерь: кросс-энтропийной и квадратичной.

Таким образом, из итоговых графиков на рис. 3 и табл. 1 видно, что предложенный алгоритм повысил качество классификации. Для рецепторов SR-HSE, SR-p53, NR-PPAR-gamma и NR-Aromatase качество увеличилось на 8%–12%, для NR-AhR, NR-AR-LBD, SR-MMP, NR-ER, SR-ARE и NR-ER-LBD — на 4%–7%, для SR-AR-LBD и SR-ATAD5 —

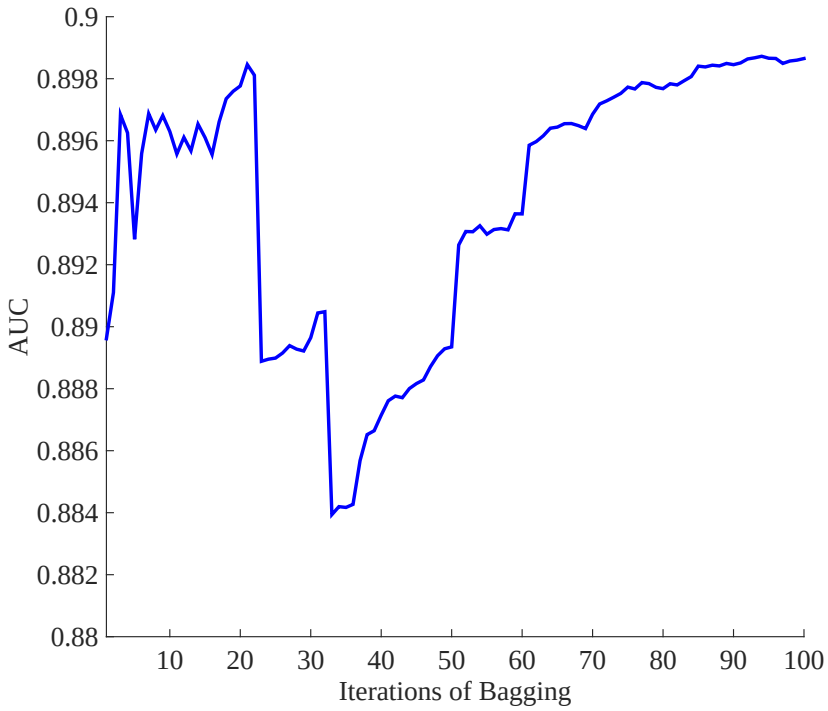

Рис. 5 График зависимости значения функционала AUC от количества разбиений на выборки в бэггинге

Таблица 1 Сравнение значений функционала AUC базового алгоритма нейронной сети и предложенного алгоритма бэггинга нейронных сетей с двумя функциями потерь: кросс-энтропийной и квадратичной

Рецептор	Нейронная сеть (кросс-энтропия)	Бэггинг (кросс-энтропия)	Нейронная сеть (квадратичная)	Бэггинг (квадратичная)
NR-AhR	0,8589 ± 0,0216	**0,9089 ± 0,0210**	0,8584 ± 0,0150	**0,9088 ± 0,0174**
NR-AR-LBD	0,8725 ± 0,0455	**0,9138 ± 0,0064**	0,9008 ± 0,0490	**0,9207 ± 0,0458**
NR-AR	0,8456 ± 0,0294	**0,8658 ± 0,0129**	0,8457 ± 0,0312	**0,8704 ± 0,0166**
SR-MMP	0,8512 ± 0,0483	**0,9132 ± 0,0110**	0,8651 ± 0,0080	**0,9161 ± 0,0109**
NR-ER	0,7585 ± 0,0726	**0,8109 ± 0,0329**	0,7545 ± 0,0414	**0,8151 ± 0,0253**
SR-HSE	0,7189 ± 0,0583	**0,8274 ± 0,0193**	0,7541 ± 0,0176	**0,8380 ± 0,0347**
SR-p53	0,7345 ± 0,0838	**0,8532 ± 0,0257**	0,7660 ± 0,0236	**0,8585 ± 0,0204**
NR-PPAR-gamma	0,7610 ± 0,0725	**0,8435 ± 0,0437**	0,7818 ± 0,0285	**0,8539 ± 0,0171**
SR-ARE	0,7698 ± 0,0307	**0,8265 ± 0,0208**	0,7652 ± 00309	**0,8268 ± 0,0076**
NR-Aromatase	0,7808 ± 0,0482	**0,8697 ± 0,0308**	0,8466 ± 0,0531	**0,8676 ± 0,0218**
SR-ATAD5	0,8338 ± 0,0714	**0,8682 ± 0,0187**	0,7713 ± 0,0648	**0,8629 ± 0,0332**
NR-ER-LBD	0,8299 ± 0,0241	**0,8917 ± 0,0267**	0,8515 ± 0,0251	**0,8884 ± 0,0168**

на 2% и 3% соответственно. Сравнивая результаты, полученные с разными функциями потерь, получаем, что значения AUC для одних и тех же рецепторов различаются максимум на 1,1% у рецептора SR-HSE, что меньше средней погрешности результатов.

Проведен вычислительный эксперимент для бэггинга, мощности подвыборок которого меньше мощности исходной выборки. Точность результатов падает с уменьшением размера подвыборки.

5 Заключение

В работе решалась проблема предсказания лиганд-рецепторного взаимодействия. В качестве модели классификации была предложена композиция двухслойных нейронных сетей — бэггинг. Рассмотрены задачи линейной и логистической регрессии с квадратичной и кросс-энтропийной функциями потерь соответственно. Исследовано изменение качества классификации с помощью декомпозиции функции ошибки на смещение и дисперсию. Проведено сравнение моделей с помощью вычислительного эксперимента на реальных данных. Полученные результаты говорят о том, что бэггинг позволяет повысить качество классификации.

Авторы выражают благодарность В. В. Стрижову за постановку задачи и внимательное отношение к работе.

Литература

[1] *Perkins R., Fang H., Tong W., Welsh W. J.* Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology // Environ. Toxicol. Chem., 2003. Vol. 22. No. 8. P. 1666–1679.

[2] *Bhasin M., Raghava G. P. S.* Classification of nuclear receptors based on amino acid composition and dipeptide composition // J. Biol. Chem., 2004. Vol. 279. No. 22. P. 23262–23266.

[3] *Salum L. B., Andricopulo A. D.* Fragment-based QSAR: Perspectives in drug design // Mol. Divers., 2009. Vol. 13. No. 3. P. 277–285.

[4] *Myint K. Z., Xie X. Q.* Recent advances in fragment-based QSAR and multi-dimensional QSAR methods // Int. J. Mol. Sci., 2010. Vol. 11. No. 10. P. 3846–3866.

[5] *Brown R. D., Martin Y. C.* The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding // J. Chem. Inf. Comp. Sci., 1997. Vol. 37. No. 1. P. 1–9.

[6] *DiMasi J. A., Hansen R. W., Grabowski H. G.* The price of innovation: New estimates of drug development costs // J. Health Econ., 2003. Vol. 22. No. 2. P. 151–185.

[7] *Zhang L., Zhu H., Oprea T. I., Golbraikh A., Tropsha A.* QSAR modeling of the blood-brain barrier permeability for diverse organic compounds // Pharm. Res., 2008. Vol. 25. No. 8. P. 1902–1914.

[8] *Myint K. Z., Wang L., Tong Q., Xie X. Q.* Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions // Mol. Pharm., 2012. Vol. 9. No. 10. P. 2912–2923.

[9] *Zhang L., Fourches D., Sedykh A., et al.* Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening // J. Chem. Inf. Model., 2013. Vol. 53. No. 2. P. 475–492.

[10] *Enrique Sucar L., Bielza C., Morales E. F., Hernandez-Leal P., Zaragoza J. H., Larrañaga P.* Multi-label classification with Bayesian network-based chain classifiers // Pattern Recogn. Lett., 2014. Vol. 41. No. 1. P. 14–22.

[11] *Popova M.* Feature selection and multi-task prediction of biological activity for nuclear receptors, technical report. URL: `https://goo.gl/5nXQMZ`.

[12] *Barkoula N. M., Alcock B., Cabrera N. O., Peijs T.* Fatigue properties of highly oriented polypropylene tapes and all-polypropylene composites // Polym. Polym. Compos., 2008. Vol. 16. No. 2. P. 101–113.

[13] *Steffen C., Thomas K., Huniar U., Hellweg A., Rubner O., Schroer A.* TmoleX — a graphical user interface for Turbomole. // J. Comput. Chem., 2010. Vol. 31. No. 16. P. 2967–2970.

[14] *Fang J., Yang R., Gao L., et al.* Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery // Mol. Divers., 2015. Vol. 19. No. 1. P. 149–162.

[15] *Gonzalez-Diaz H., Bonet I., Teran C., et al.* ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds // Eur. J. Med. Chem., 2007. Vol. 42. No. 5. P. 580–585.

[16] *Patra J. C., Chua K. H. K.* Neural network based drug design for diabetes mellitus using QSAR with 2D and 3D descriptors // Joint Conference (International) on Neural Networks Proceedings, 2010. P. 18–23.

[17] *Tu J. V.* Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes // J. Clin. Epidemiol., 1996. Vol. 49. No. 11. P. 1225–1231.

[18] *Lisboa P.* A review of evidence of health benefit from artificial neural networks in medical intervention // Neural Networks, 2002. Vol. 15. No. 1. P. 11–39.

[19] *Ha K., Cho S., Maclachlan D.* Response models based on bagging neural networks // J. Interact. Mark., 2005. Vol. 19. No. 1. P. 17–30.

[20] *Zhou Z. H., Wu J., Tang W.* Ensembling neural networks: Many could be better than all // Artif. Intell., 2002. Vol. 137. No. 1-2. P. 239–263.

[21] *Tibshirani R.* Bias, variance and prediction error for classification rules. University of Toronto, Department of Statistics, 1996.

[22] *James G. M.* Variance and bias for general loss functions // Mach. Learn., 2001. Vol. 51. No. 2. P. 115–135.

[23] *Geman S., Bienenstock E., Doursat R.* Neural networks and the bias/variance dilemma. — 1992. P. 1–58.

Поступила в редакцию 15.09.2016

Bagging of neural networks for analysis of nuclear receptor biological activity*

M. R. Vladimirova and M. S. Popova

mrvladimirova@gmail.com; popova@gmail.com

Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, Russia

The paper is devoted to the multitask classification problem. The main purpose is building an adequate model to predict whether the object belongs to a particular class, precisely, whether the ligand binds to a specific nuclear receptor. Nuclear receptors are a class of proteins found within cells. These receptors work with other proteins to regulate the expression of specific genes, thereby controlling the development, homeostasis, and metabolism of the organism. The regulation of gene expression generally only happens when a ligand — a molecule that effects the receptor's behavior — binds to a nuclear receptor. Two-layer neural network is used as a classification model. The paper considers the problems of linear and logistic regressions with squared and cross-entropy loss functions. To analyze the classification result, the authors propose to decompose the error into bias and variance terms. To improve the quality of classification by reducing the error variance, they suggest the composition of neural networks: the bagging procedure. The proposed method improves the quality of the investigated sample classification.

Keywords: *nuclear receptors; biological activity; two-layer neural network; bagging; multitask learning; drug-design; cross-entropy*

DOI: 10.21469/22233792.2.3.06

*This research is funded by the Russian Foundation for Basic Research, grant 16-07-01155.

References

[1] Perkins, R., H. Fang, W. Tong, and W. J. Welsh. 2003. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* 22(8):1666–1679.

[2] Bhasin, M., and G. P. S. Raghava. 2004. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279(22):23262–23266.

[3] Salum, L. B., and A. D. Andricopulo. 2009. Fragment-based QSAR: Perspectives in drug design. *Mol. Divers.* 13(3):277–285.

[4] Myint, K. Z., and X. Q. Xie. 2010. Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *Int. J. Mol. Sci.* 11(10):3846–3866.

[5] Brown, R. D., and Y. C. Martin. 1997. The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comp. Sci.* 37(1):1–9.

[6] DiMasi, J. A., R. W. Hansen, and H. G. Grabowski. 2003. The price of innovation: New estimates of drug development costs. *J. Health Econ.* 22(2):151–185.

[7] Zhang, L., H. Zhu, T. I. Oprea, A. Golbraikh, and A. Tropsha. 2008. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* 25(8):1902–1914.

[8] Myint, K. Z., L. Wang, Q. Tong, and X. Q. Xie. 2012. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol. Pharm.* 9(10):2912–2923.

[9] Zhang, L., D. Fourches, A. Sedykh, *et al.* 2013. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J. Chem. Inf. Model.* 53(2):475–492.

[10] Enrique Sucar, L., C. Bielza, E. F. Morales, P. Hernandez-Leal, J. H. Zaragoza, and P. Larrañaga. 2014. Multi-label classification with Bayesian network-based chain classifiers. *Pattern Recogn. Lett.* 41(1):14–22.

[11] Popova, M. *Feature selection and multi-task prediction of biological activity for nuclear receptors, technical report.* Available at: `https://goo.gl/5nXQMZ` (accessed December 20, 2015).

[12] Barkoula, N. M., B. Alcock, N. O. Cabrera, and T. Peijs. 2008. Fatigue properties of highly oriented polypropylene tapes and all-polypropylene composites. *Polym. Polym. Compos.* 16(2):101–113.

[13] Steffen, C., K. Thomas, U. Huniar, A. Hellweg, O. Rubner, and A. Schroer. 2010. TmoleX — a graphical user interface for Turbomole. *J. Comput. Chem.* 31(16):2967–2970.

[14] Fang, J., R. Yang, L. Gao, *et al.* 2015. Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery. *Mol. Divers.* 19(1):149–162.

[15] Gonzalez-Diaz, H., I. Bonet, C. Teran, *et al.* 2007. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* 42(5):580–585.

[16] Patra, J. C., and K. H. K. Chua. 2010. Neural network based drug design for diabetes mellitus using QSAR with 2D and 3D descriptors. *Joint Conference (International) on Neural Networks Proceedings.* 18–23.

[17] Tu, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* 49(11):1225–1231.

[18] Lisboa, P. 2002. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks* 15(1):11–39.

[19] Ha, K., S. Cho, and D. Maclachlan. 2005. Response models based on bagging neural networks. *J. Interact. Mark.* 19(1):17–30.

[20] Zhou, Z. H., J. Wu, and W. Tang. 2002. Ensembling neural networks: Many could be better than all. *Artif. Intell.* 137(1-2):239–263.

[21] Tibshirani, R. 1996. Bias, variance and prediction error for classification rules. University of Toronto, Department of Statistics.

[22] James, G. M. 2001. Variance and bias for general loss functions. *Mach. Learn.* 51(2):115–135.

[23] Geman, S., E. Bienenstock, and R. Doursat. 1992. Neural networks and the bias/variance dilemma. 1–58.

Received September 15, 2016