Применение нейронных сетей для автоматизации экологического мониторинга цианобактериальных «цветений» водоемов

Н. Ю. Григорьева, Л. В. Чистякова
Ресурсный центр «Культивирование микроорганизмов»
Научный парк
Санкт-Петербургский государственный университет
renes3@mail.ru

А. А. Лисс, Д. М. Клионский, А. С. Перков, Т. Р. Жангиров
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
anna.liss@moevm.info

***Аннотация*. В данной работе исследуется возможность применения нейронных сетей в задачах оперативной классификации цианобактерий при непрерывном экологическом мониторинге водных объектов. Предложена инновационная методика дифференциации цианобактерий на основе совместного применения новейших методов конфокальной микроспектроскопии и нейросетевых технологий обработки данных.**

Ключевые слова: нейронные сети; задача классификации; экологический мониторинг

I. Ведение

Одна из задач экологического мониторинга водных объектов – это исследование и контроль токсичных цианобактериальных «цветений». Данная задача включает в себя два аспекта: исследование биологического разнообразия цианобактерий в водоеме и выявление и непрерывный контроль развития потенциально опасных штаммов. До настоящего времени эта задача была трудно формализуема вследствие неоднозначности и высокой трудоемкости методов сбора информации. Однако недавно авторами данной статьи были разработаны две уникальные методики: методика определения таксономической принадлежности цианобактерий [1] и методика оценки жизнеспособности цианобактериальных культур [2] на основе данных флуоресцентной микроспектроскопии отдельных клеток.

Спектры собственной флуоресценции клеток цианобатерий, снятые на конфокальном микроскопе [3], удобны для статистической обработки и последующего анализа различными математическими методами. Кроме того, сам процесс сбора первичных данных флуоресцентного анализа может быть легко автоматизирован на базе on-a-chip технологии [4]. Таким образом, данное исследование в совокупности с ранее разработанными методиками может быть положено в основу создания программно-аппаратного комплекса для непрерывного экологического мониторинга открытых водоемов.

В данной работе представлена одна из возможных реализаций задачи определения таксономической принадлежности цианобактерий на основе нейронной сети прямого распространения, а также проведено детальное исследование нескольких методов обучения при различных параметрах сети и количестве исходных признаков, используемых для классификации. Для рассматриваемой задачи классификации предложена структура нейронной сети, оптимальная с точки зрения соотношения «точность классификации-время обучения». Проведено сравнение результатов классификации, полученных в рамках использования искусственной нейронной сети и линейного дискриминантного анализа.

II. Описание исходных данных

Исходные экспериментальные данные представляют собой серии из семи спектров собственной флуоресценции, полученные на конфокальном лазерном сканирующем микроскопе Leica TCS-SP5. Спектры снимались при различных длинах волн возбуждающего излучения, соответствующих лазерным линиям 405, 458, 476, 488, 496, 514, 543 нм. В дальнейшем для краткости наборы признаков, выделенных из каждого спектра серии будут нумероваться по длине волны соответствующей лазерной линии. Каждый исходный спектр представляет собой массив из 38–45 чисел, соответствующих интенсивностям флуоресценции на определенных частотах излучения видимого диапазона от 520 до 785 нм.

Для извлечения из исходных данных набора классификационных признаков в математическом пакете MATLAB была разработана программа. В рамках данной программы осуществляется интерполяция, экстраполяция и сглаживание исходных спектров, приведение их к общему масштабу и размеру массива, взятие первой производной и проведение быстрого Фурье-преобразования, а также расчет конкретных величин, характеризующих форму кривых и спектрального состава их производных. После дополнительного исследования был определен набор из 63 признаков, область значений которых лежит в диапазоне [-1,1]. Все признаки можно

разделить на три группы: асимметрия и эксцесс (АЕ), процентное соотношение флуоресценции отдельных пигментов по четырем основным областям спектра (ПСФ) и частотные характеристики соответствующих Фурье-трансформант для каждого графика (процентное соотношение вкладов по трем частотным областям) (ПФТ).

В ходе данной работы было использовано 314 наборов из 7 спектров, соответствующих 21 штамму и 15 родам цианобактерий из коллекции *CALU* Ресурсного центра «Культивирование микроорганизмов» НП СПбГУ.

III. Проектирование нейронной сети

Вследствие простоты решаемой задачи изначально рассматривалась трехслойная нейронная сеть (НС) прямого распространения с одним скрытым слоем. В качестве функции активации на скрытом слое использовался гиперболический тангенс, а на выходном слое – функция Softmax. На обоих слоях присутствует нейрон сдвига с сигналом равным 1. В ходе исследования рассматривались несколько вариантов сетей с разным количеством входных нейронов (N_{in}), зависящих от количества признаков, используемых для классификации. Количество нейронов на выходном слое было фиксировано по количеству классов ($N_{out}=16$). Количество нейронов на скрытом слое оценивалось по формуле $N_h \sim \sqrt{N_{in}N_{out}}$ [5].

Для решения задачи классификации в качестве алгоритмов обучения НС использовались методы обратного распространения ошибки. Были рассмотрены 8 градиентных методов оптимизации первого порядка: GD – метод градиентного спуска; AdaGrad – метод адаптивного градиента; AdaDelta – метод адаптивного шага обучения; Adam – метод оценки адаптивного момента; CG – метод сопряжённых градиентов; NAG – метод ускоренного градиента Нестерова; QProp – метод быстрого обратного распространения; RProp – метод упругого обратного распространения. Данные методы отличаются способом корректировки весов и минимизации функции ошибки.

Так как на выходном слое используется активационная функция Softmax, то для расчета ошибки выхода сети применяется функция кросс-энтропии.

IV. Результаты и обсуждение

В ходе данной работы были исследованы зависимости качества классификации данных нейронной сетью от: метода обучения, скорости обучения, количества эпох обучения, числа нейронов на скрытом слое, способа инициализации начальных весов. Кроме того, в процессе моделирования НС использовались результаты ступенчатого линейного дискриминантного анализа (ЛДА) при определении оптимальных наборов классификационных признаков, а также проводилась оценка корректности результатов работы НС на основе сравнения с ЛДА. Вследствие ограниченности объема данной публикации приведем только несколько основных результатов.

A. *Выбор метода обучения*

На одном и том же наборе данных различные методы обучения могут оптимизировать НС по-разному. Поэтому необходимо рассмотреть несколько методов и выбрать оптимальный для конкретной классификационной задачи. Основными критериями отбора в данном случае являются сходимость и устойчивость метода обучения. На рис. 1(a) приведена зависимость ошибки выхода НС от количества эпох для различных методов обучения. Во всех расчетах соотношение обучающей и тестовой выборки было 70:30 %. Другие параметры модели были следующие: порог допустимой ошибки обучения – 0.01, размер окна контроля ошибки – 20, параметр момента – 0.1, параметр регуляризации – 0.001.

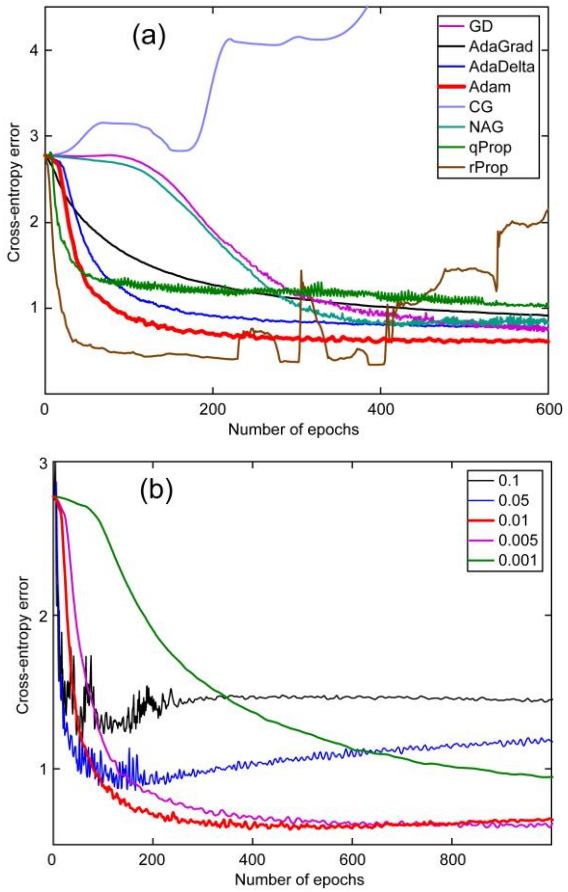

Рис. 1. Изменение ошибки от количествава эпох: (a) – для различных методов обучения при скорости обучения 0.01, (b) – для метода Adam при различных скоростях обучения.

Согласно рис. 1(a) минимальное значение ошибки наблюдается для методов Adam и rProp, но для последнего по достижении минимума ошибка начинает резко возрастать и становится нестабильной. Поэтому для дальнейших расчетов был выбран метод обучения Adam.

Поскольку выбор метода обучения проводился при фиксированной скорости обучения, которая влияет на

скорость минимизации ошибки, для выбранного метода Adam было проведено исследование зависимости ошибки выхода нейронной сети от количества эпох при различных скоростях обучения. Из представленных на рис. 1(b) зависимостей следует, что при скорости обучения больше 0.05, ошибка не достигает минимума и начинает осциллировать и возрастать, что свидетельствует о том, что она уходит из области глобального минимума. При скоростях 0.005–0.01 значение ошибки доходит до предполагаемого глобального минимума и находится в устойчивом состоянии при числе эпох от 300 до 800. Проводить обучение при скоростях меньших 0.005 нет смысла, так как при этом уменьшение ошибки происходит крайне медленно. Таким образом, выбранная скорость обучения 0.01 является оптимальной для данной задачи.

B. Сравнение результатов ЛДА и НС

В данной работе ЛДА использовался на нескольких этапах исследования. На этапе отбора наиболее значимых признаков классификации ступенчатый ЛДА использовался для формирования нескольких наборов, отличающихся по количеству признаков и точности классификации. На конечном этапе ЛДА применялся также и для оценки результатов классификации.

ТАБЛИЦА I СРАВНЕНИЕ РЕЗУЛЬТАТОВ КЛАССИФИКАЦИИ ЛДА И НС

Набор признаков (количество)	***ЛДА***	***НС***
1. АЭ (405, 458, 488, 543) **(8)**	80.9 %	79.1 %
2. ПСФ (405, 458, 488, 543) **(16)**	93.5 %	78.4 %
3. АЭ, ПСФ (405, 458, 488, 543) **(24)**	93.7 %	85.6 %
4. АЭ, ПФТ (405, 458, 488, 543) **(21)**	93.6 %	91.1 %
5. АЭ, ПСФ и ПФТ (405, 458, 488, 543) **(36)**	97.3 %	93.8 %
6. ПСФ[a] **(28)**	95,1%	79.7 %
7. АЭ (405, 458, 488, 543) и ПСФ[a] **(42)**	97,5%	90.0 %
8. АЭ (405, 458, 488, 543), ПСФ[a] и ПФТ[a] **(57)**	98,5%	95.4 %

[a] *Полный набор лазерных линий (405, 458, 476, 488, 496, 514, 543)*

В таблице приведены значения точности классификации ЛДА и НС для 8 наборов признаков. Использовалась выборка из 236 наблюдений, приписанных к 16 классам. Результаты получены на сети с 30 нейронами на скрытом слое и на 300 эпохах при скорости обучения 0.01. Первые 5 наборов классификационных признаков были сформированы на основе результатов ступенчатого ЛДА. Ограниченный набор лазерных линий (4 вместо 7) рассматривался с учетом возможной аппаратной реализации блока сбора первичной информации. Поскольку для разработки микроэлектронного прибора на основе on-a-chip технологии необходимо ограничить число источников возбуждения флуоресценции и количество спектральных зон приема излучения. Для этой цели было проведено сравнение точности классификации для полного и ограниченного набора лазерных линий.

Более высокая точность классификации для ЛДА обуславливается тем, что он работает с функциями распределения признаков и их статистическими характеристиками, что позволяет лучше построить модель классификатора. Но, при этом, для ЛДА на признаки накладываются более сильные ограничения по наличию корреляций. Кроме того, преимуществом НС является возможность продолжения обучения модели без необходимости заново строить классификатор при появлении новых экспериментальных данных. Возможность дообучать НС, дает потенциальную возможность повысить точность ее результатов дополняя исходную выборку новыми наблюдениями. Заметим, что при проведении ЛДА рассмотрение большого количества признаков может быть затруднено чисто технически, в этом случае НС может стать единственной альтернативой.

C. Проверка качества обобщения НС

В поставленной задаче классификации качество работы НС должно определяться не только абсолютным значением точности классификации, но и способностью построенной сети распознавать и правильно классифицировать новые данные, не принадлежащие ни к одному из классов, участвовавших в обучении.

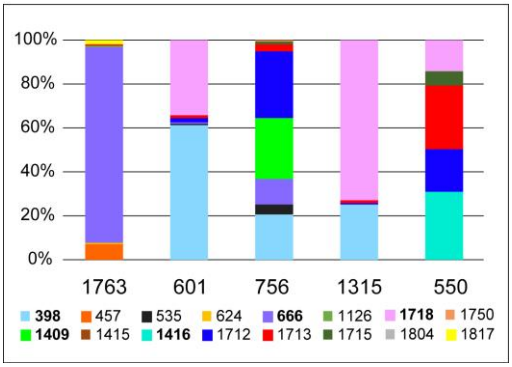

Рис. 2. Результаты классификации 5 новых штаммов, не входивших в исходную выборку из 16 штаммов. Цифры соответствуют номерам штаммов в коллекции *CALU*

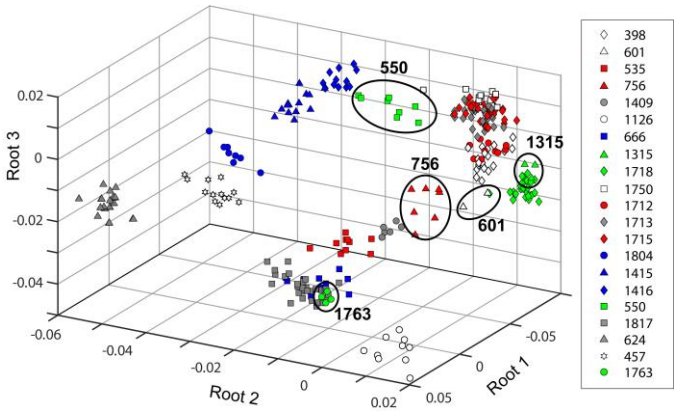

Рис. 3. Результаты классификации ЛДА для 21 штамма. Цифры соответствуют номерам штаммов в коллекции *CALU*

На рис. 2. представлены результаты классификации 5 новых штаммов, не входивших в исходную выборку. Для расчетов использовался набор параметров 7 из таблицы. Классификатор НС должен был определить, к каким из 16 известных классов можно было бы отнести 5 неизвестных штаммов. Близкородственными штаммами в данном случае являлись 1763-666, 601-398, 756-1409, 1315-1718,

550-1416 (в парах первый штамм не определен для НС, а второй – из обучающей выборки). Небольшие погрешности в классификации штаммов 756 и 550 можно объяснить тем, что данные штаммы в пространстве параметров лежат в промежутках между 16 тестовыми штаммами, примерно на равных расстояниях от 2–3 ближайших. Поэтому НС не смогла выбрать корректное решение. Это наглядно демонстрирует рис. 3, на котором представлены результаты ЛДА Фишера для полной выборки из 21 штамма. Замкнутые линии показывают области, занятые пятью тестовыми штаммами.

V. Заключение

Таким образом, в данной работе была спроектирована нейронная сеть для задачи классификации цианобактерий по родам и штаммам. Проведен анализ различных методов обучения и выбран метод Adam, как наиболее оптимальный по соотношению «точность классификации-время обучения», а также как показавший максимальную устойчивость на предложенном наборе признаков. Были оптимизированы такие параметры нейронной сети как количество нейронов на скрытом слое, скорость обучения, количество эпох необходимых для минимизации ошибки. Для валидации корректности работы нейронной сети было проведено сравнение результатов классификации НС с результатами ЛДА, а также обученная НС была протестирована на качество распознавания новых классов данных, не представленных в обучающей и тестовой выборке. Кроме того, была рассмотрена возможность аппаратной реализации всего комплекса сбора и обработки информации на базе on-a-chip технологии. Сравнение результатов анализа для полного и ограниченного набора классификационных признаков показало незначительное снижение точности классификации.

Список литературы

[1] Флуоресцентная микроспектроскопия для исследования биологического разнообразия цианобактерий в пресноводных экосистемах / Л.В. Чистякова, Т.Р. Жангиров, А.А. Лисс, Н.Ю. Григорьева // Биоиндикация в мониторинге пресноводных экосистем: Материалы III междунар. конф., СПб 16-20 окт. 2017 / под. ред. В.А. Румянцева, И.С. Трифоновой. СПб: Свое Издательство, 2017. С. 365-369.

[2] Румянцев В.А., Григорьева Н.Ю, Чистякова Л.В. Исследование изменений физиологического состояния цианобактерий после слабого ультразвукового воздействия // Доклады Академии наук. 2017. Т. 475, № 5. С. 580-583.

[3] Pawley J.B. Handbook of Biological Confocal Microscopy. NY, London: Plenum Press, 1995. 600 p.

[4] Harrison D.J., Fluri K., Seiler K., Fan Z., Effenhauser C.S., Manz A. Micromachining a miniaturized capillary electrophoresis-based chemical analysis system on a chip. // Science. 1993. V. 261 (5123). P. 895-897.

[5] Хайкин С.С. Нейронные сети: полный курс. М: Издательский дом «Вильям», 2006. 1104 с.