ИНЖИНИРИНГ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 631 .67: 621.647.2: 621.643: 621.67

О НЕКОТОРЫХ АСПЕКТАХ ИСПОЛЬЗОВАНИЯ НЕЙРОННЫХ СЕТЕЙ ПРИ РЕШЕНИИ ЗАДАЧ БИОИНФОРМАТИКИ

Дунай Валерий Иванович, к.б.н., доцент
Штепа Владимир Николаевич, д.т.н., доцент
Глинская Наталья Анатольевна, к.с.-х.н., доцент
Полесский государственный университет
Dunay Valeriy Ivanovich, PhD., tppoless@gmail.com
Shtepa Vladimir Nikolayevich, Dr.,
Glinskaya Nataliya Anatolyevna, PhD
Polessky State University

Рассмотрена возможность использования искусственных нейронных сетей для решения практических задач в биоинформатике. Предложен базовый алгоритм использования нейросетей в задачах, без учёта итерационных действий, при определении расположения белков. Сформулировано направление дальнейших исследований применения искусственного интеллекта в биоинформатике.

Ключевые слова: *биоинформатика, нейронная сеть, фильтрация сигнала, расположение белка, байесовские сети.*

С развитием информационных технологий (ИТ) появилось множество прорывных решений разного характера. На фоне такого прогресса ИТ в биологии сформировался и нашёл применение новый раздел – биоинформатика (БИ) [1].

Её специфика в том, что она оперирует не непосредственно живыми объектами, а результатами их исследований. То есть объектом изысканий может быть полный геном организма, отдельный ген, белковые взаимодействия, эволюционные исследования; но между специалистами с биоинформатики и объектом исследования стоит экспериментатор, который эти данные добыл, после которого такой материал нужно интерпретировать. Но возникают прецеденты, когда БИ используется с самого начала исследований, еще на стадии планирования, чтобы изначально знать, как полученные данные будут обрабатываться и прогнозировать конечный результат.

Биоинформатика находит применение в различных областях. В первую очередь, конечно, в медицине [2]. Сейчас уже понятно, что многие болезни имеют генетическую природу, то есть либо заложены в геноме человека уже с рождения, либо развиваются в результате возникающих в нем мутаций. Классический пример – онкологические заболевания. Известно, что клетка, прежде чем стать раковой, претерпевает множество геномных изменений. Ее ДНК изменяется, куски хромосом меняются местами. Эти процессы очень гетерогенны: у разных пациентов с одним и тем же типом опухоли раковые клетки могут мутировать по-разному, что затрудняет борьбу с такими болезнями.

ДНК раковых клеток расшифровываются по всему миру, и сейчас уже есть открытые базы, из которых можно скачивать геномные данные опухолей, сравнивать с тем, что наблюдается у конкретного больного, и лучше понимать, как его лечить. Более того, анализ значительного количества геномов раковых клеток, в том числе с помощью машинного обучения, позволил выявить сильно мутирующие очаги в геноме, на которые нужно смотреть в первую очередь. В 2018 году ученые из американского Университета Джона Хопкинса опубликовали статью о возможности создать на основе накопленных знаний о мутациях панель диагностики рака на ранних стадиях, когда сам человек еще ничего не чувствует и сканирование ничего не показывает. А в анализах крови уже можно найти мутантную ДНК, что является непосредственной задачей биоинформатики.

Аналогично можно диагностировать предрасположенность и ко многим другим опасным болезням, но для этого необходимо использование специализированных математических аппаратов, которые потом реализовываются в прикладных программных обеспечениях (ПО).

К таким интеллектуальным решениям относятся искусственные нейронные сети (НС) – класс моделей, построенных с использованием алгоритмов машинного обучения на основе принципа о том, что мыслительные явления могут быть описаны сетями из взаимосвязанных простых элементов, по аналогии с организацией биологических нейронных сетей. НС имитирует поведение системы, исходя из предоставленных экспериментальных или известных из других источников данных, позволяя пропустить этап создания алгоритмической/механической модели, необходимый для описания системы и решения связанных с ней задач при традиционном подходе и представляющий значительные трудности для сложных и нелинейных систем, часто встречающихся в задачах из области биологии [3].

Так как определение расположения белков экспериментальными методами требует больших временных затрат и является дорогостоящим, а механизмы сортировки достаточно хорошо изучены только для небольшого количества белков и их возможных локализаций, в условиях поступления все большего количества данных, полученных в результате секвенирования, автоматические методы решения этой задачи становятся все более востребованными [4]. Одним из рассматриваемых сегодня в научной литературе методов является использование характеристик белков, таких как аминокислотная последовательность, дипептидный состав и др., в качестве опорных точек для автоматических систем предсказания [5].

Вместе с тем практическое использование НС для определения расположения белков, как показывает практика [6], возможно только с применением предпроцессирования – предварительной обработки входной информации, например, математической фильтрации сигналов. Соответственно, базовый алгоритм использования НС в задачах БИ может включать ряд этапов, без учёта итерационных действий:

1. формирование набора экспериментальных данных;
2. математическая фильтрация информационной составляющей входных в НС сигналов;
3. синтез и параметрирование НС;
4. оценка работы НС с биологической точки зрения, внесение структурных и функциональных корректировок;
5. штатное использования НС, например, в задачах определения расположения белков.

Более детально остановимся на пунктах 2 и 3 такой последовательности. Для фильтрации результатов экспериментальных исследований обосновано использовать преобразование Гильберта-Хуанга, которое оперирует методом эмпирической модовой декомпозиции (EMD). Он базируется на предположении, что любой набор данных содержит различные режимы колебательных процессов. Каждый из таких колебательных режимов может быть представлен функцией внутренней моды (IMF) с соответствующими ограничениями:

- количество экстремумов и количество нулевых сечений функции должны быть равными или отличаться не более чем на единицу;
- в любой точке функции среднее значение огибающих кривых, определенных локальными экстремумами, должно быть равно 0.

То есть IMF представляют собой колебательные режимы, которые вместо постоянных амплитуды и частоты могут иметь переменные амплитуду и частоту как функции времени.

Суть EMD заключается в последовательном (итерационном) установлении функций эмпирических мод cj(t) и остатков rj(t) = rj-1(t) – cj(t), где j = 1, 2, 3, …, n при r0 = y(t). Результатом разложения будет представление сигнала в виде суммы модовых функций и конечного остатка [7]:

$$x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t), \quad (1)$$

где n — количество эмпирических мод, устанавливаемое при расчете.

Исследования [7] продемонстрировали, что соответствующий адаптивный базис хотя и не определен аналитически, но удовлетворяет требованиям традиционных базисов: завершенности, сходимости, ортогональности и единственности (утверждение спорное).

При синтезе непосредственно НС определения расположения белков, пункт 3 предложенной последовательности, целесообразно использовать архитектуру вероятностных нейронных сетей [8], построенных на вероятностных моделях, представляющих собой множество переменных и их вероятностных зависимостей Байесовской статистики. Формально байесовская сеть – это направленный ациклический граф, вершинами которого являются переменные, а ребра кодируют условные зависимости между переменными. Вершины могут представлять переменные любых типов, быть взвешенными параметрами, скрытыми переменными или гипотезами. Если ребро выходит из вершины A в вершину B, то A называют отцом B, а B называют потомком A. Множество вершин предков вершины Xi обозначим как parents (Xi), тогда общее распределение значений в вершинах можно удобно расписать как результат локальных распределений:

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i)). \quad (2)$$

где n – количество локальных распределений.

К частному случаю байесовских сетей относятся вероятностные нейронные сети (probnabilistic neural networks – PNN) – вид нейронных сетей, эффективно применяемых для решения задач классификации, где плотность вероятности принадлежности классам оценивается с помощью ядерной аппроксимации. При решении задач классификации выходы сети можно с пользой интерпретировать как оценку вероятности или элемент принадлежит некоторому классу. Сеть практически учится оценивать функцию плотности вероятности. Аналогичная полезная интерпретация может возникать и в задачах регрессионного анализа – выход сети рассматривается как ожидаемое значение модели в конкретной точке пространства входов. Это ожидаемое значение связано с плотностью вероятности совместного распределения входных и выходных данных.

В основу классификации в сети PNN положена идея, что для каждого образца можно принять решение на основе выбора наиболее вероятного класса из тех, которым мог бы принадлежать этот образец. Такое решение требует оценки функции плотности вероятностей для каждого. Эта оценка устанавливается в результате рассмотрения обучающих данных. Правилом является то, что образ с плотным распределением в области неизвестного образца будет значительным по сравнению с другими элементами. Точно так же будет иметь преимущество и образ с высокой априорной вероятностью или высокой ценой ошибки классификации. Для двух классов A и B, согласно этому правилу, выбирается класс A, если:

$$h_A c_A f_A(x) = h_B c_B f_B(x), \quad (3)$$

где h – априорная вероятность; c – цена ошибки классификации; f(x) – функцию плотности вероятностей.

Правильная оценка ошибки классификации требует точного знания предметной области (биоинформатики), но во многих случаях она и априорная вероятность выбираются одинаковыми для всех классов. Оценить функции плотности распределения вероятностей можно с помощью метода Парцена, в котором используется весовая функция, имеющая центр в точке, представляющей обучающий образец. Такая весовая функция называется потенциальной функцией или ядром. Чаще всего в качестве ядра используется функция Гаусса. Чтобы построить функцию распределения имущественности, для каждого вектора рассматривается функция Гаусса с центром в точке, соответствующей этому вектору. Затем эти функции суммируются, после чего получается разыскиваемая функция распределения. Традиционно используется более простая форма функции Гаусса, где включен квадрат евклидового расстояния от неизвестного образа до элемента слоя образцовых образов:

$$g(x) = \sum_{i=1}^{n} \exp\left(\frac{-\left\|x - x_i\right\|^2}{\sigma^2}\right). \quad (4)$$

Соответственно, комбинирование математической фильтрации входных сигналов и вероятностных нейронных сетей позволит качественно определять не только расположения белков, но и решать другие прикладные задачи биоинформатики.

Дальнейшие исследования обосновано нацелить на формирование базы данных результатов экспериментов и создание специализированного ПО с использованием математического аппарата НС и преобразования Гильберта-Хуанга, как вариант на основе высокоуровневого языка программирования общего назначения Python.

Выводы. Нейросетевые методы совместно с предпроцесированием позволят анализировать значимые массивы данных биологического характера. В тоже время современные информационные технологии хоть и не являются абсолютным научным инструментарием, но переводят многие исследования на совершенно новый уровень по скорости и точности исполнения, что открывает значимые перспективы для развития биоинформатики.

Список использованных источников

1. Афонников, Д.А. Биоинформатика: метод во главе угла / Д.А. Афонников, В.А. Иванисенко // Наука из первых рук, 2013. – Т.49, N 1. – С.50–59.

2. Арчаков, А.И. Биоинформатика, геномика и протеомика – науки о жизни XXI столетия / А.И. Арчаков // Вопросы медицинской химии, 2000. – Т.46, N 1. – С.4–7.

3. Horton, P. A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins / P. Horton, K. Nakai // In: International Conference on Intelligent Systems for Molecular Biology. – St. Louis: AAAI Press, 1996. –P. 109 –115.

4. LOCATE: a mammalian protein subcellular localization database / J. Sprenger [et all.] // Nucleic Acids Res, 2008. – V. 36. – P. D230–D233. doi: 10.1093/nar/gkm950.

5. Al-mubaid, H. New Feature Weighting Technique for Predicting Protein Subcellular Localization / H. Al-mubaid, D.B. Nguyen. – In: 2014 IEEE International Conference on Bioinformatics and Bioengineering. – Boca Raton: IEEE, 2014. doi: 10.1109/BIBE.2014.35.

6. Mellman, I. Coordinated protein sorting, targeting and distribution in polarized cells / I. Mellman, W. J. Nelson // Nature Reviews Molecular Cell Biology, 2008. – V. 11. – № 11. – P. 833–845. doi: 10.1038/nrm2525.

7. Intelligent effective management system of biotechnical objects based on natural disturbances prediction / V. Lysenko [et all.] // Earth Bioresources and Life Quality, 2013. – № 4. – P. 34–41.

8. Вероятностные нейронные сети в задачах управления комбинированными системами водоочистки / В.Н. Штепа [и др.] // Вестник Брестского государственного технического университета. Сер. Водохозяйственное строительство, теплоэнергетика и геоэкология: научно теоретический журнал. – 2018. – № 2 (110). – С. 88–90.