ЕСТЕСТВЕННЫЕ НАУКИ

УДК 004.622, 616.006

АЛГОРИТМ ОБРАБОТКИ ДАННЫХ РАМАН-СПЕКТРОСКОПИИ ДЛЯ ФОРМИРОВАНИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ НЕЙРОННОЙ СЕТИ В ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКЕ ЗЛОКАЧЕСТВЕННЫХ ЗАБОЛЕВАНИЙ

Ю. А. Субхангулова [1], А. Т. Бикмеев [2], А. Р. Билялов [2]

[1] 42nd.marvin@gmail.com

[1, 2] ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)
[3] ФГБОУ ВО «Башкирский государственный медицинский университет» (БГМУ)

Аннотация. Работа посвящена разработке алгоритма построения обучающей выборки для обучения искусственной нейронной сети в условиях ограниченности как объема данных, так и времени принятия решения. Предложен метод снижения размерности исходных данных на основе вейвлет-разложения, а также определены его оптимальные параметры, такие как тип используемых вейвлетов и уровень разложения. В результате удалось сократить время обучения нейронной сети с нескольких часов до десятков секунд, при этом снижение точности классификации тканей было незначительным, оно составило менее 0,5%.

Ключевые слова: вейвлет; раман-спектроскопия; обработка сигналов; онкология; нейронная сеть; обучающая выборка; обработка данных.

ВВЕДЕНИЕ

Известно, что в среднем около четверти онкологических больных (26%) умирают в течение года после постановки диагноза [1]. Важнейшую роль играет стадия заболевания, так как именно от нее зависит степень распространения опухолевого процесса и выбор методов терапии. Наиболее хорошо поддаются лечению 1 и 2 стадия заболевания, прогнозы пятилетней выживаемости после проведения адекватной терапии в эти периоды составляют 93% и 75%. На 4 стадии пятилетняя выживаемость равна 15%, иногда – не более 5%. Таким образом, своевременная и максимально точная диагностика онкологических заболеваний, желательно малоинвазивным способом, является важнейшей задачей.

Прогнозы выживаемости в ходе лечения зависят от множества факторов: возраст пациентов, вид злокачественного новообразования, сопутствующие патологии и других. Большое влияние на процесс выздоровления после оперативного вмешательства имеет качество проведенной операции с точки зрения удаления всех злокачественных образований, дабы не допустить рецидива заболевания. Следовательно, хирургу-онкологу необходимы средства оперативной диагностики качества тканей человека в ходе хирургической операции, что накладывается на такие инструменты достаточно серьезное требование получения результата в максимально короткие сроки: секунды и минуты.

Следует отметить, что раковые клетки отличаются по химическому составу и структуре от здоровых клеток. Это приводит к идее использования спектроскопических методов анализа тканей для выявления очагов поражения как на ранних стадиях (поскольку изменения в концентрации химических веществ и структуре тканей начинаются задолго до явного проявления болезни), так и в ходе оперативного вмешательства.

Такие работы проводятся как за рубежом [2, 3], так и у нас в России [4, 5]. В указанных работах для анализа тканей использовалась спектроскопия комбинационного рассеяния (Раман-спектроскопия). В последних двух из них анализ проводился с использованием разработанной нейронной сети, при этом размерность исходных данных понижалась путем разложения спектрограмм по вейвлетам Добеши. Однако авторы не представили детального обоснования выбора метода подготовки данных и не исследовали применимость более сложных вейвлетов для анализа Раман-спектрограмм.

В связи с вышесказанным, представляет интерес систематический и последовательный анализ методики и построение алгоритма подготовки спектрограмм для формирования обучающей выборки для нейронной сети.

ОПИСАНИЕ ЗАДАЧИ

Типичный вид спектрограммы Раман-спектроскопии представлен на рис. 1. Видно, что он представляет собой кривую довольно сложной формы, на которую наложен шум.

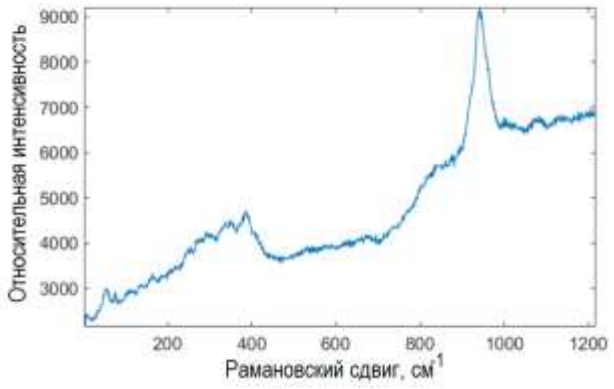

Рис. 1. Типичный вид спектрограммы Раман-сектроскопии

В работе [4] показано, что использование спектрограмм тканей одного и того же органа одного пациента позволяет достичь наилучшей точности в диагносте заболевания. Однако такой подход не позволяет получить выборку большого объема. В результате задача обучения нейронной сети по данным спектроскопии одного пациента имеет несколько особенностей:

− ограниченный набор спектрограмм, ввиду небольшого набора образцов;

− большой объем исходных данных – каждая спектрограмма в исходном виде насчитывает более 1000 точек;

− зашумленность исходных данных, что связано как с точностью прибора, так и с разбросом в химическом составе тканей человека.

Разрабатываемая методика должна учесть все эти особенности и, по возможности, устранить их негативное влияние на результат диагностики.

АЛГОРИТМ ПОДГОТОВКИ ОБУЧАЮЩЕЙ ВЫБОРКИ

Поскольку идентификация химического состава, а значит и определение принадлежности тканей к тому или иному типу, выполняется на основе анализа расположения и высоты пиков на спектрограмме, то, как предложено в работе [5], в начале проведем выравнивание спектра по оси абсцисс, то есть вычтем из него, так называемую базовую линию. Затем, при помощи функций Wavelet Toolbox пакета Matlab, удалим шум. Результат такой обработки можно увидеть на рис. 2.

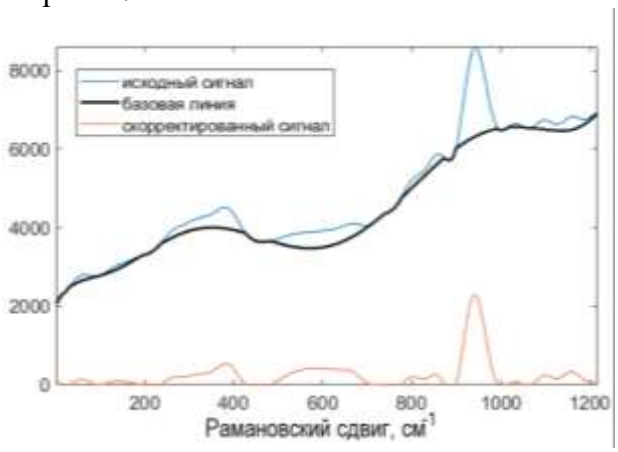

Рис. 2. Результат сглаживания и вычитания базовой линии

Прежде всего рассмотрим возможность использования спектрограмм в исходном виде. Для этого будем использовать полносвязную нейронную сеть (НС) следующей топологии: 1217 элементов входного слоя, 2500 нейронов на первом скрытом слое, 100 нейронов на втором скрытом слое и 2 нейрона на выходе. Выходной вектор ин-

терпретируется следующим образом: (0,1) соответствует здоровой ткани, (1,0) – пораженной. В качестве функции активации нейронов используем логический сигмоид. Обучение НС проводится методом обратного распространения ошибки.

Ввиду ограниченного набора данных (20 спектрограмм) обучающая выборка только на основе имеющихся данных вряд ли позволит получить качественный результат. Поэтому проведем искусственное «размножение» исходных данных. Если представить спектрограмму в виде:

$$y = s + b + \varepsilon,$$

где y – спектрограмма, s – полезный сигнал, b – базовая линия, ε – шум, то, предположив, что шум является белым гауссовым и определив его характеристики, можно сформировать любое количество искусственных сигналов вида:

$$y' = s + b + \varepsilon'$$

где ε' – сгенерированный белый шум.

В результате для искусственно расширенной выборки из трехсот спектрограмм время обучения для 25 эпох (ошибка стабилизируется на 23 эпохе) составляет около 1 часа при точности идентификации по тестовой выборке 96,7%.

Для уменьшения размерности данных проведем исследование эффективности разложения спектрограмм по вейвлетам до 10-го уровня: на первом уровне разложения спектрограмма содержит 609 точек, а на 10-м – 2 точки. В качестве основы разложения будем использовать следующие вейвлеты: вейвлет Добеши, симлет и биортогональный вейвлет.

В соответствии с количеством точек, в разложении меняется число входов НС, а количество нейронов двух скрытых слоев вычисляется согласно теореме Колмогорова [6].

В результате проведения большой серии экспериментов определен наилучший вариант разложения: вейвлет Добеши 3-го порядка с глубиной разложения 5-го уровня. Итоговый вектор содержит всего 39 точек. Обучение НС на 300 векторах сокращенной размерности (25 эпох) занимает около 10 секунд, а точность предсказания по тестовой выборке составляет 96,32%, то есть отличие от точности обучения по исходной спектрограмме составляет менее 0,5%.

Таким образом, алгоритм обработки данных Раман-спектроскопии с целью получения обучающей выборки для диагностирующей нейронной сети может быть записан следующим образом: 1) определить дисперсию шума спектрограмм, 2) удалить из спектрограмм шум и базовую линию, 3) провести искусственное размножение спектрограмм до нескольких сотен, 4) выполнить разложение спектрограмм до уровня 5 по вейвлетам Добеши 3-го порядка.

СПИСОК ЛИТЕРАТУРЫ

1. **Каприн А.Д.** Злокачественные новообразования в России в 2017 году (заболеваемость и смертность). М.: МНИОИ им. П.А. Герцена – филиал ФГБУ «НМИЦ радиологии» Минздрава России, 2018. 250 с. **[**A.D. Kaprin, Malignant neoplasms in Russia in 2017 (morbidity and mortality) (in Russian). Moscow: Moscow Cancer Research Institute named after P.A. Herzen – branch of the Federal State Budgetary Institution Scientific Research Center for Radiology, Ministry of Health of Russia, 2018. 250 p.**]**

2. **Li Q.B., Wang W., Liu Ch.-H., Zhanga G.-J.** Discrimination of breast cancer from normal tissue with Raman spectroscopy and chemometrics // Journal of Applied Spectroscopy. 2015. Т. 82. С. 450–455. **[**Q.B. Li, W. Wang, Ch.-H. Liu, G.-J. Zhanga, “Discrimination of breast cancer from normal tissue with Raman spectroscopy and chemometrics” in Journal of Applied Spectroscopy, vol. 82, pp. 450-455, 2015.**]**

3. **Jermyn M., Desroches J., Aubertin K.** A review of Raman spectroscopy advances with an emphasis on clinical translation challenges in oncology // Phys. Med. Biol. 2016. Т. 61, № 23. С. R370–R400. [M. Jermyn, J. Desroches, K. Aubertin, “A review of Raman spectroscopy advances with an emphasis on clinical translation challenges in oncology” in Phys. Med. Biol., vol. 61, no. 23, pp. R370-R400 2016.]

4. **Павлов В.Н., Билялов А.Р., Гильманова Р.Ф.** Использование интеллектуальных методов обработки данных раман-спектроскопии для диагностики злокачественных опухолей // Медицинский вестник Башкортостана. 2018. Т. 13, № 3 (75). С. 43–47. **[**V.N. Pavlov, A.R. Bilyalov, R.F. Gilmanova, “The use of intelligent data processing techniques of raman-spectroscopy for the diagnosis of malignant tumors”, (in Russian), in Meditsinskiy vestnik Bashkortostana, vol. 13, no. 3 (75), pp. 43-47, 2018.**]**

5. **Павлов В.Н., Билялов А.Р., Ковтуненко А.С.** Использование метода комбинационного рассеяния света при диагностике опухолевых заболеваний человека // Актуальные вопросы биологической физики и химии. 2018. Т. 3, № 4. С.874–879. **[**V.N. Pavlov, A.R Bilyalov., A.S. Kovtunenko, “Using the method of combination light scattering in diagnosis of human”, (in Russian), in Aktualnyye voprosy biologicheskoy fiziki i khimii, vol. 3, no. 4, pp. 874-879, 2018.**]**

6. **Колмогоров А.Н.** О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения. Докл. АН СССР. 1957. Т. 114, № 5. С.953–956. **[**A. N. Kolmogorov, “On the representation of continuous functions of several variables as superpositions of continuous functions of one variable and addition”, (in Russian), in Dokl. USSR Academy of Sciences, vol. 114, no. 5, pp. 953-956, 1957.**]**

ОБ АВТОРАХ

СУБХАНГУЛОВА Юлия Анваровна, магистрант 1-го курса общенаучного факультета ФГБОУ ВО «УГАТУ».

БИКМЕЕВ Александр Тимерзянович, канд. физ.-мат. наук, зав. кафедрой цифровых технологий в нефтяном инжиниринге ФГБОУ ВО «УГАТУ».

БИЛЯЛОВ Азат Ринатович, канд. мед. наук, начальник управления информационных технологий ФГБОУ ВО БГМУ Минздрава России.

METADATA

Title: Web OLAP conceptual data model design on the basis of situation-oriented database

Authors: J. A. Subkhangulova [1], A. T. Bikmeyev [2], A. R. Bilyalov [3]

Affiliation:

[1, 2] Ufa State Aviation Technical University (UGATU), Russia.

[3] Bashkortostan State Medical University (BSMU), Russia.

Email: [1] 42nd.marvin@gmail.ru

Language: Russian.

Source: Molodezhnyj Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), no. 2 (23), pp. 135-138, 2020. ISSN 2225-9309 (Print).

Abstract: The work is devoted to the development of an algorithm for making a training dataset for an artificial neural network in conditions of limited data volume and decision time. A method for reducing the dimension of the initial data based on the wavelet decomposition is proposed, and its optimal parameters are determined, such as the type of wavelets used and the level of decomposition. As a result, it was possible to reduce the training time of the neural network from hours to seconds, while the decrease in the accuracy of the classification of tissues was not significant, it was less than 0.5%.

Key words: wavelet; Raman spectroscopy; signal processing; oncology; artificial neural network; train dataset; data processing.

About authors:

SUBKHANGULOVA, Julia Anvarovna, Bachelor of Science, Ufa State Aviation Technical University.

BIKMEYEV, Alexandr Timerzyanovich, Ph. D. in phisico-matematical scince, Ufa State Aviation Technical University.

BILYALOV, Azat Rinatovich, Ph. D. in medical scince, Bashkortostan State Medical University.

ОПЕЧАТКИ

Стр.	Напечатано	Следует читать
138 (название статьи на английском)	**Title:** Web OLAP conceptual data model design on the basis of situation-oriented database	**Title:** An algorithm of Raman spectroscopy data processing to generate a training dataset for an artificial neural network in the differential diagnosis of malignant diseases
148 (формула (5))	$m\frac{S_{i,j}^{n+1}-S_{i,j}^{n}}{\tau}-$ $-\frac{k}{\mu_B}\frac{1}{h}\left[k_B\left(S_{i+\frac{1}{2},j}^{n+1}\right)\frac{p_{i+1,j}^{n+1}-p_{i,j}^{n+1}}{h}-\right.$ $-kBSi-12,$ $jn+1pi,jn+1-pi-1,jn+1h++kBSi,$ $j+12n+1pi,j+1n+1-pi,jn+1h--kBSi$ $,j-12npi,jn+1-pi,j-1n+1h=0,$	$m\frac{S_{i,j}^{n+1}-S_{i,j}^{n}}{\tau}-$ $-\frac{k}{\mu_B}\frac{1}{h}\left[k_B\left(S_{i+\frac{1}{2},j}^{n+1}\right)\frac{p_{i+1,j}^{n+1}-p_{i,j}^{n+1}}{h}-\right.$ $-k_B\left(S_{i-\frac{1}{2},j}^{n+1}\right)\frac{p_{i,j}^{n+1}-p_{i-1,j}^{n+1}}{h}+$ $+k_B\left(S_{i,j+\frac{1}{2}}^{n+1}\right)\frac{p_{i,j+1}^{n+1}-p_{i,j}^{n+1}}{h}-$ $\left.-k_B\left(S_{i,j-\frac{1}{2}}^{n}\right)\frac{p_{i,j}^{n+1}-p_{i,j-1}^{n+1}}{h}\right]=0,$
147-150 (статья Абдрахманова А.Р. «Об использовании техники автоматического дифференцирования на примере решения задачи двумерной фильтрации»)	отнесена к разделу «Гуманитарные науки»	относится к разделу «Технические науки»