КОНФЕРЕНЦИЯ А

МОЛЕКУЛЯРНАЯ СПЕКТРОСКОПИЯ И АТМОСФЕРНЫЕ РАДИАЦИОННЫЕ ПРОЦЕССЫ

КЛАССИФИКАЦИЯ БОЛЬНЫХ БРОНХО-ЛЕГОЧНЫМИ ЗАБОЛЕВАНИЯМИ НА ОСНОВЕ АНАЛИЗА СПЕКТРОВ ПОГЛОЩЕНИЯ ПРОБ ВЫДЫХАЕМОГО ВОЗДУХА С ПРИМЕНЕНИЕМ МЕТОДА ОПОРНЫХ ВЕКТОРОВ, НЕЙРОННОЙ СЕТИ

Кистенев Ю.В.[1,2], Кузьмин Д.А.[2], Вражнов Д.А.[3], Борисов А.В. [1,2]

[1]ФГАОУ ВО НИ ТГУ, Российская Федерация, 634050, г. Томск, пр. Ленина, 36

[2] ГБОУ ВПО СибГМУ Минздрава России, г.Томск

[3] ООО «Томсклаб», 634055, г. Томск, пр-т Академический, д. 8/8

E-mail: yuk@iao.ru, band107@mail.ru, denis.vrazhnov@gmail.com, borisov@phys.tsu.ru

Ключевые слова: выдыхаемый воздух, лазерная оптико-акустическая спектроскопия, метод главных компонент, метод опорных векторов, рак легких, хроническая обструктивная болезнь легких, нейронная сеть

Аннотация

В работе представлены результаты классификации больных бронхо-легочными заболеваниями на основе анализа проб выдыхаемого воздуха, полученных с помощью применения метода лазерной оптико-акустической спектроскопии и методов интеллектуального анализа данных (метод главных компонент, метод опорных векторов, нейронная сеть). Были зарегистрированы спектры поглощения выдыхаемого воздуха набранных добровольцев, проведена подготовка данных к процедуре классификации спектров поглощения выдыхаемого воздуха больных и здоровых людей, а также определены матрицы ошибок в случае нейронной сети и чувствительность/специфичность в случае метода опорных векторов для полученных результатов классификации.

Работа выполнена при частичной финансовой поддержке ФЦП ИР, контракт №14.578.21.0082 (уникальный идентификатор прикладных научных исследований и экспериментальных разработок RFMEFI57814X0082).

В работе рассматривается задача классификации спектров проб выдыхаемого воздуха (ПВВ), полученных с помощью метода лазерной оптико-акустической спектроскопии (ЛОАГ).

Ранее было показано [1], что в задаче классификации различных групп пациентов возможен подход анализа спектров поглощения ПВВ групп без решения обратной спектроскопической задачи и без выделения определенных газовых компонент. В работе [2] получены данные применения метода опорных векторов (SVM-метод, от англ. «support vector machine») для попарной групповой классификации некоторых нозологических состояний (рак легкого, хроническая обструктивная болезнь легких, пневмония в сравнении с условным состоянием здоровья). Представляет интерес сравнение возможностей SVM-метода и метода нейронных сетей в задаче классификации ПВВ.

В данной работе, аналогично [2] исследовались 3 группы добровольцев общим количеством 30 человек, количество участников в каждой группе одинаково.

В первую группу (группа 1) вошли больные с верифицированным диагнозом рак легкого (РЛ). Локализация, степень развития патологического процесса были различными. Все больные РЛ являлись пациентами, проходившими обследование в торакоабдоминальном отделении ФГБУ НИИ онкологии СО РАМН (ФГБНУ Томский национальный исследовательский медицинский центр РАН в г. Томске). Численность – 10 человек. Средний возраст в группе 1 – 56,4 года. Критерии исключения: неверифицированный диагноз, прохождение лечения (химиотерапия, радиотерапия, хирургическое лечение), тяжелое течение сопутствующих заболеваний, наличие других бронхо-легочных заболеваний.

Во вторую группу (группа 2) вошли больные с верифицированным диагнозом хроническая обструктивная болезнь легких (ХОБЛ) в фазе обострения. Степень тяжести заболевания была различной. Все больные ХОБЛ являлись пациентами, проходившими обследование в пульмонологическом отделении ОГАУЗ Городская клиническая больница №3 (г. Томск). Численность – 10 человек. Средний возраст в группе 2 – 53,1 года. Критерии исключения: неверифицированный диагноз, наличие других бронхо-легочных заболеваний, тяжелое течение сопутствующих заболеваний.

В третью группу (группа 3) вошли условно-здоровые добровольцы, некурящие. Критерии включения: отсутствие острых заболеваний в течение 2 недель, предшествующих забору проб, отсутствие хронических заболеваний бронхо-легочной, пищеварительной, сердечно-сосудистой и мочеполовой систем, преимущественное отсутствие фактора курения в анамнезе. Численность – 10 человек. Средний возраст в группе 3 – 24,7 года.

Воздух собирался в стандартную пробирку объемом 10 мл. Доброволец выполнял несколько обычных выдохов через пластиковую трубочку непосредственно в пробирку, закрывающуюся плотным стерильным ватным тампоном. Все пробы отбирались в утреннее время, до еды или через 2 часа после нее. Курящие испытуемые не курили до забора проб хотя бы в течение 30 минут. До взятия проб испытуемые полоскали ротовую полость проточной водой.

Спектры поглощения проб регистрировались с использованием лазерных оптико-акустических газоанализаторов ILPA-1 и ЛГА-2, разработанных фирмой ООО «Специальные технологии» (г. Новосибирск). Лазерные газоанализаторы ILPA-1, ЛГА-2 собраны на базе волноводных, перестраиваемых по частоте в диапазоне 9,2-10,8 мкм CO_2-лазеров и резонансных оптико-акустических детекторов. Конструктивные особенности: ILPA-1 имеет внутрирезонаторное, ЛГА-2 – внерезонаторное расположение детекторов.

Для устранения экспериментальных данных с выбросами измерений использовалась методика, основанная на критерии Граббса [3]. Также была проведена процедура

интеркалибровки сканов, зарегистрированных на разных приборах. Суммарный объем выборки составлял 260 сканов.

В качестве алгоритма классификации был использован метод опорных векторов. Основной идеей метода является перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве.

Перед реализацией классификации проводился этап обучения SVM-метода. Для этого каждая группа исследования случайным образом была поделена на 2 равные выборки, одна из которых использовалась для обучения метода (обучающая выборка), а вторая (тестовая выборка) – для классификации. Обучающая выборка использовалась для построения разделяющей гиперплоскости между исследуемыми группами. После этого проводилась классификация тестовой выборки. Классификация проведена попарно, и в каждом случае найдена чувствительность и специфичность [4].

Для этих данных был применен SVM-метод для попарной групповой классификации добровольцев по их нозологическому состоянию. Полученные результаты представлены в таблице 1.

Таблица 1. Полученные результаты чувствительности и специфичности SVM-метода для попарной классификации исследуемых групп

Попарная групповая классификация	Чувствительность	Специфичность
РЛ – Здоровые	100%	63,75-67,5%*
ХОБЛ – Здоровые	95-98,75%*	92,5-93,75%*
РЛ – ХОБЛ	100%	97,5-98,75%*

Метод SVM обеспечивает бинарную классификацию, поэтому является актуальным использование классификаторов, позволяющих разделение более чем на две группы. В качестве такого классификатора в данной работе используется нейронная сеть, а также проводится сравнение возможностей классификации с помощью комбинации метода опорных векторов совместно с методом главных компонент (МГК/SVM).

Нейронные сети (НС) как инструмент классификации данных возникли как альтернатива известным условным методам классификации. Преимущества НС определяются тем, что они могут адаптироваться к данным без какой-либо их спецификации, а также они могут аппроксимировать любую функцию с любой точностью. НС преобразуют данные нелинейно, что делает их гибким инструментом моделирования сложных реальных данных и их преобразований. Кроме того, НС могут оценивать апостериорные вероятности и могут

использоваться для создания правил статистической классификации медицинских данных для медицинской диагностики. Нейронные сети могут состоять из произвольного количества нейронов, сгруппированных в один или несколько слоев. Выбор количества нейронов и конфигурации сети зависит от конкретной задачи и может варьироваться от десятков до десятков тысяч в задачах классификации спектров [5]. Изначально, наиболее популярными были нейронные сети на основе самообучающихся карт Кохонена и НС, обученные по правилу обратного распространения ошибки [6]. В настоящее время, огромный интерес вызывают сверточные нейронные сети, но их применение ограничивается областью распознавания образов.

Основным недостатком нейронных сетей является отсутствие гарантированного положительного результата при обучении, то есть, нет явного указания, как нужно выбирать конфигурационные параметры нейронной сети, чтобы получить хороший классификатор. Однако сложности при обучении компенсируются эффективностью классификации.

В работе использовалась двухслойная нейронная сеть прямого распространения с одним скрытым слоем. Особенностью данной нейронной сети является использование различных функций активаций для каждого слоя. В рассматриваемом случае применялась сигмоидальная функция активации на скрытом слое и функция активации softmax на выходном слое. Функция softmax представляет собой нормированную экспоненту:

$$y(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{k} e^{s_j}}, \quad i = 1, \ldots, k$$

где k – размерность входного вектора, s_i – вектор входных данных.

Особенность данной функции в том, что сумма выходных значений равна единице, а также частная производная i-го нейрона по своему сумматору равна

$$\frac{\partial y_i}{\partial s_i} = y_i(1 - y_i).$$

Описанная выше конфигурация нейронной сети выбрана потому, что при наличии достаточного количества нейронов в скрытом слое она обладает хорошей обобщающей способностью.

Для обучения нейронной сети использовался масштабированный метод сопряженных градиентов [7]. Преимущество данного метода в скорости работы: она на порядок выше, чем у метода обратного распространения ошибки.

На рисунке 1 показана общая схема конфигурации используемой нейронной сети. В качестве входных данных использовались спектры поглощения проб выдыхаемого воздуха,

взятые у пациентов с раком легких (РЛ), ХОБЛ и здоровых пациентов. На первом этапе, в каждом классе создавалась псевдослучайная последовательность индексов, соответствующих изучаемым спектрам. Далее, 20 спектров отбирались для итогового тестирования обученной нейронной сети, остальные использовались для обучения, валидации и первичного тестирования. На втором этапе, тестовые данные подавались на вход нейронной сети, где 35% из них было использовано для обучения, 15% для валидации и 50% для первичного тестирования. На третьем этапе происходило тестирование обученной нейронной сети при помощи 60 тестовых примеров, по 20 примеров от каждого класса. Для визуализации результата были построены матрицы ошибок. Следует отметить, что качество обучения нейронной сети зависит от начальных данных, задаваемых случайным образом. Поэтому, зачастую требуется несколько раз переобучить нейронную сеть для достижения приемлемого результата.

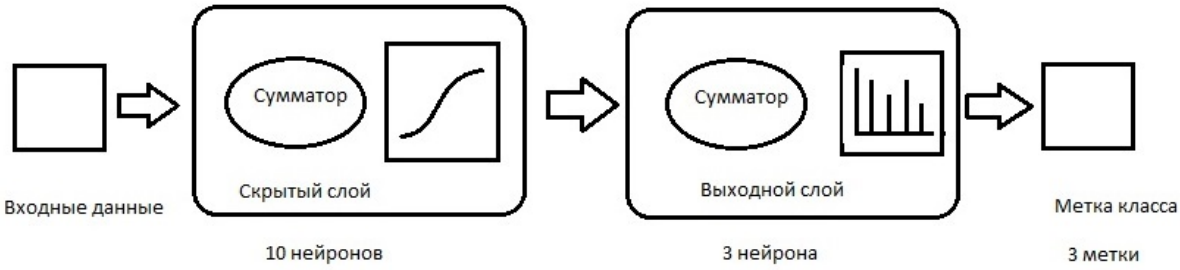

Рисунок 1. Общая схема конфигурации использованной нейронной сети

Ниже, в таблицах 2-5 приводятся матрицы ошибок для нейронной сети. В процентах указаны доли от общего числа классифицируемых спектров. На пересечении строки и столбца матрицы, приводится количество спектров из класса, предсказанного классификатором отнесенного к классу, определенному экспертом. То есть, на пересечении строки ХОБЛ и столбца РЛ таблицы 2 стоит 0, это значит, что 0 спектров из класса ХОБЛ были отнесены классификатором к классу РЛ.

Таблица 2 – Матрица ошибок тренировочной базы.

		Класс, определенный экспертом			
		РЛ	ХОБЛ	Здоровые	Итого
Класс, предсказанный классификатором	РЛ	37 37.4%	0 0.0%	0 0.0%	100%
	ХОБЛ	0 0.0%	34 34.3%	0 0.0%	100%
	Здоровые	0 0.0%	0 0.0%	28 28.3%	100%
	Итого	100%	100%	100%	100%

Таблица 3 – Матрица ошибок при валидации.

		Класс, определенный экспертом			
		РЛ	ХОБЛ	Здоровые	Итого
Класс, предсказанный классификатором	РЛ	7 23.3%	0 0.0%	0 0.0%	100%
	ХОБЛ	0 0.0%	14 46.7%	0 0.0%	100%
	Здоровые	0 0.0%	0 0.0%	9 30%	100%
	Итого	100%	100%	100%	100%

Основным показателем качества обучения служит первичное тестирование. Именно матрица ошибок (Таблица 4) первоначального тестирования отражает то, насколько хорошо обученная нейронная сеть обобщает входные данные и не совершает ошибочных классификаций.

Таблица 4 – Матрица ошибок при первичном тестировании.

		Класс, определенный экспертом			
		РЛ	ХОБЛ	Здоровые	Итого
Класс, предсказанный классификатором	РЛ	20 29%	0 0.0%	1 1.4%	95.2%
	ХОБЛ	0 0.0%	34 34.3%	0 0.0%	100%
	Здоровые	0 0.0%	0 0.0%	24 34.8%	100%
	Итого	100%	100%	96%	98.6%

В таблице 5 представлен суммарный результат классификации таблиц 2-4.

Таблица 5 – Итоговая матрица ошибок.

		Класс, определенный экспертом			
		РЛ	ХОБЛ	Здоровые	Итого
Класс, предсказанный классификатором	РЛ	64 32.3%	0 0.0%	1 0.5%	98.5%
	ХОБЛ	0 0.0%	72 36.4%	0 0.0%	100%
	Здоровые	0 0.0%	0 0.0%	61 30.8%	100%
	Итого	100%	100%	98.4%	99.5%

Качество работы обученного классификатора определяется на тестовой выборке. Дополнительное вторичное тестирование с предварительно случайным образом выбранными спектрами используется как дополнительная независимая проверка (Таблица 6).

Таблица 6 – Матрица ошибок при вторичном тестировании.

		Класс, определенный экспертом			
		РЛ	ХОБЛ	Здоровые	Итого
Класс, предсказанный классификатором	РЛ	19 31.7%	0 0.0%	1 1.7%	95.0%
	ХОБЛ	0 0.0%	20 33.3%	0 0.0%	100%
	Здоровые	1 1.7%	0 0.0%	19 31.7%	95.0%
	Итого	95.0%	100%	95.0%	96.7%

Для сравнения качества классификации с помощью МГК/SVM [8-9] и нейронной сети, составим таблицу специфичности и чувствительности [10] для МГК/SVM на данных, используемых при классификации с помощью нейронной сети.

Идея применения МГК/SVM состоит в следующем: ко всем спектрам ПВВ, применяется МГК и с помощью SVM проводится бинарная классификация главных компонент, сравнивая каждую из главных компонент одного множества с главной компонентой другого множества.

В таблице 7 показано применение МГК/SVM для классификации трех групп по следующему принципу: с помощью классификатора три раза производится разделение двух множеств, одно из которых является спектрами ПВВ искомой группы, а второе состоит из смеси двух оставшихся групп. Таким образом, имея три варианта обучающей выборки для классификатора SVM можно осуществить разделение на группы – РЛ, ХОБЛ, Здоровые.

Для каждой пары произведено усреднение характеристик TPR (false-positive rate - чувствительность) и FPR (true-positive rate - специфичность) по 50 различным обучающим выборкам. При этом расчеты проведены для следующих ядер: Linear, Quadratic, Polynomial, Gaussian RadialBasis Function, Multilayer Perceptron (mlp) kernel с различными параметрами и для попарных главных компонент от 1 до 10. В таблице 7 показан наиболее качественный результат.

Из сравнения таблиц 1 и 7 очевидно, что в точность разделения бинарным классификатором SVM в случае двух множеств выше. Однако, из полученных результатов следует, что группы РЛ, ХОБЛ, Здоровые обладают существенными различиями.

Таблица 7 – Пример применения МГК/SVM.

	Вариант 1		Вариант 2		Вариант 3	
	РЛ (TPR)	ХОБЛ и Здоровые (FPR)	ХОБЛ (TPR)	РЛ и Здоровые (FPR)	Здоровые (TPR)	РЛ и ХОБЛ (FPR)
Среднее	0.8271	0.9481	0.9792	0.9854	0.9999	0.9122
Дисперсия	0.0785	0.0969	0.0730	0.0928	0.0575	0.0418

Обученная нейронная сеть, несмотря на малый размер обучающей выборки, показывает хорошие результаты на тестовой выборке. Проведенные тесты с повторной генерацией случайных выборок тестовых и тренировочных баз позволяет сделать вывод о наличии специфических черт, характерных для классов РЛ, ХОБЛ, Здоровые, различие которых обучена находить нейронная сеть.

Таким образом, классификации с помощью МГК/SVM и нейронной сети приводят к аналогичным результатам.

Работа выполнена при частичной финансовой поддержке ФЦП ИР, контракт №14.578.21.0082 (уникальный идентификатор прикладных научных исследований и экспериментальных разработок RFMEFI57814X0082).

Литература

1. E.B. Bukreeva; A.A. Bulanova; Y.V. Kistenev et al. Analysis of the absorption spectra of gas emission of patients with lung cancer and chronic obstructive pulmonary disease by laser optoacoustic spectroscopy - SPIE Proceedings Vol. 8699, 2013.

2. E.B. Bukreeva; A.A. Bulanova; Y.V. Kistenev et al. Application of support vector machine method for the analysis of absorption spectra of exhaled air of patients with broncho-pulmonary diseases SPIE Proceedings Vol. 92926 2014.

3. Frank E. Grubbs. Procedures for Detecting Outlying Observations in Samples // Technometrics, 1969. – Vol. 11. – No. 1. – P.1-21.

4. Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. – Cambridge University Press, 2000.

5. Gasteiger J. Zupan J. Neural Networks for Chemists: An Introduction; 1st – s.l.: VCH, 1993.

6. Kohonen T. Self-Organization and Associative Memory; 8 – s.l.: Springer Berlin Heidelberg, 1989.

7. Moller M. F., A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. Neural Networks, vol. 6(4), pp. 525 533, 1993.

8. Célia Lourenço and Claire Turner. Breath Analysis in Disease Diagnosis: Methodological Considerations and Applications// Metabolites 2014, 4, 465-498.

9. Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers // Advances in Kernel Methods. - MIT Press, Cambridge, USA, 1998.

10. Powers, David M W Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation // Journal of Machine Learning Technologies 2 (1): 37–63, 2011.