УДК 616.2

ПРИМЕНЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ ИНТЕРПРЕТАЦИИ НАРУШЕНИЙ МЕХАНИКИ ДЫХАНИЯ И ЛЕГОЧНОГО ГАЗООБМЕНА

А.В. Тишков[1], М.Ю. Каменева[2], А.А. Гладской[1], А.И. Гунченко[1], В.И. Трофимов[2]

[1]Санкт-Петербургский институт информатики и автоматизации РАН
[2]Санкт-Петербургский государственный медицинский университет им. И.П.Павлова

Предложен способ выявления закономерностей в показателях спирометрии и бодиплетизмографии, диффузионной способности и растяжимости легких у больных идиопатическим фиброзирующим авльвеолитом и гистиоцитозом Х легких на основе метода классификации при помощи деревьев решений. Показано, что для дифференциальной диагностики идиопатического фиброзирующего альвеолита и гистиоцитоза Х легких необходимо комплексное применение легочных функциональных тестов.

Ключевые слова: *идиопатический фиброзирующий авльвеолит, гистиоцитоз Х легких, механика дыхания, газообмен легких, раскопка данных, деревья решений.*

Введение. В современной пульмонологии методы функционального исследования внешнего дыхания являются неотъемлемой частью протоколов обследования и лечения больных. Оснащение современных клиник и учреждений амбулаторно-поликлинической помощи аппаратами для проведения легочных функциональных тестов в последние годы значительно улучшилось и большинство трудностей, с которыми сталкиваются специалисты — это вопросы стандартизации измерений и интерпретации полученных результатов. Диагностические возможности отдельных методов изучены недостаточно, а вопросам комплексного использования различных методов уделяется чрезвычайно мало внимания [2; 5].

Для понимания возможностей применения методов исследования механики дыхания и легочного газообмена в диагностике заболеваний легких могут применяться математические методы “раскопки данных” или Data Mining, которые с точки зрения авторов пока недостаточно широко используются при анализе медицинских данных, уступая место признанным статистическим подходам. В данной работе представлен пример использования одного из наиболее распространенных методов раскопки данных – деревья решений [6].

Методика. Были проанализированы две выборки пациентов, для которых известен диагноз: 28 больных идиопатическим фиброзирующим альвеолитом (ИФА) и 37 больных гистиоцитозом Х

легких (ГХЛ). Этот выбор был связан с тем, что при ИФА формируется рестриктивный тип вентиляционных нарушений, а для ГХЛ характерна весьма пестрая картина нарушений механики дыхания с преобладанием так называемого «смешанного» типа, не имеющего определенных диагностических критериев.

Построенное дерево решений позволяет отнести пациента к одному из этих двух заболеваний, иначе говоря — классифицировать пациентов на два класса. Хотя такое дерево не может устанавливать предварительный диагноз, поскольку ничего не знает о других диагнозах и здоровых пациентах, его структура интересна сама по себе и может послужить основой для гипотез и дальнейших исследований.

Всем пациентам было выполнено комплексное функциональное исследование внешнего дыхания. Анализировались следующие показатели: жизненная емкость легких вдоха (ЖЕЛ) и форсированная (ФЖЕЛ), объем форсированного выдоха за первую секунду ($ОФВ_1$), мгновенная объемная скорость при выдохе 50% ФЖЕЛ ($МОС_{50}$), средняя объемная скорость при выдохе от 25% до 75% ФЖЕЛ ($СОС_{25-75}$), общая емкость легких (ОЕЛ), остаточный объем легких (ООЛ), диффузионная способность легких (ДСЛ), соотношение ДСЛ и объема альвеолярной вентиляции (ДСЛ/АВ), динамическая растяжимость (CL) и индекс ретракции (CR).

Результаты спирометрии оценивали с помощью должных величин, разработанных Р.Ф. Клементом с соавт. [3], статические легочные объемы – по должным величинам, предложенным Европейским сообществом угля и стали [4]. Интерпретация других показатели механики дыхания проводилась соответственно рекомендациям, изложенным в руководстве «Интерстициальные заболевания легких» [1].

В табл. 1 представлены значения должных величин для 10 из перечисленных показателей и абсолютное значение для индекса ретракции.

Во втором и пятом столбцах табл. 1 указаны средние значения показателей и стандартные отклонения (СО), в третьем и шестом столбцах — 95%-ные непараметрические доверительные интервалы для среднего. Р - это значение по критерию Манна-Уитни.

Отметим, что согласно данным последнего столбца табл. 1, имеются шесть показателей, статистически значимо различающихся у пациентов с ИФА и ГХЛ, не включая возраст. Наиболее выражено различие по ООЛ. Напротив, показатели $ОФВ_1$, $СОС_{25-75}$, ДСЛ/АВ и ЖЕЛ не различаются в этих группах.

Дерево решений устроено таким образом, что в узлах располагаются показатели, разделяющие так называемую обучающую выборку на подгруппы. Обучающей выборка называется потому, что вместе с набором атрибутов-показателей, для каждого объекта

определена метка класса, в нашем случае ИФА или ГХЛ. Эта метка указывает алгоритму построения дерева какие атрибуты-показатели характеризуют класс ИФА, а какие класс ГХЛ. Ребра, ведущие из вершины, соответствуют условиям разбиения текущей группы объектов на подгруппы. Последние узлы, «листья» дерева соответствуют классу ИФА или ГХЛ.

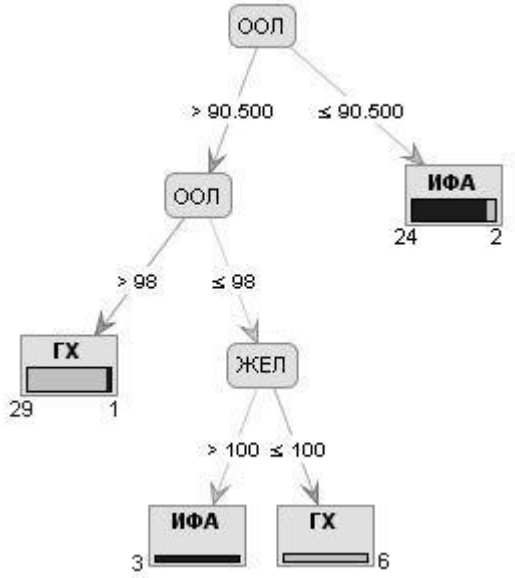
Таблица 1

Описательная статистика групп ИФА и ГХЛ

Показатели	ИФА (N=28)			ГХЛ (N=37)			
	среднее (СО)	95%ДИ	диапазон	среднее (СО)	95%ДИ	диапазон	P
Возраст, годы	50,12 (13,27)	[45,20;55,03]	[19,55;65,34]	31,87 (10,38)	[28,53; 35,21]	[16,10;57,96]	$1 \cdot 10^{-6}$ ***
ЖЕЛ, %Долж	82,46 (19,72)	[73,90;91,06]	[37;126]	87,76 (14,18)	[82,0;93,51]	[52;120]	0,4909
$ОФВ_1$, %Долж	81,86 (22,89)	[73,40;90,31]	[41;133]	81,08 (24,30)	[73,25;88,91]	[27;125]	0,9419
$МОС_{50}$, %Долж	77,94 (34,88)	[27,57;47,47]	[2,38;131]	61,24 (27,70)	[50,35;72,13]	[7;140]	0,0434 *
$СОС_{25-75}$, %Долж	50,30 (33,56)	[37,86;62,73]	[0,3;169]	52,90 (27,80)	[42,08;63,74]	[6,6;128]	0,7911
ОЕЛ, %Долж	77,14 (16,37)	[69,80;84,48]	[39;116]	97,35 (17,29)	[13,95;20,62]	[66;135]	0,0001 ***
ООЛ, %Долж	73,86 (12,20)	[67,73;79,99]	[46;125]	123,97 (37,82)	[21,89;53,76]	[56;272]	$2 \cdot 10^{-9}$ ***
ООЛ/ОЕЛ, %Долж	95,36 (16,96)	[12,39;21,52]	[74;143]	123,70 (20,95)	[114,62;132,79]	[82;222]	$1 \cdot 10^{-5}$ ***
ДСЛ, %Долж	53,36 (16,19)	[46,02;60,69]	[18;103]	58,99 (18,82)	[51,71;66,27]	[18; 104]	0,3368
ДСЛ/АВ, %Долж	71,07 (15,63)	[64,01;78,13]	[34;109]	66,51 (17,20)	[59,38;73,65]	[18; 104]	0,4993
CL, % Долж	57,64 (44,59)	[23,9;65,69]	[8;224]	88,13 (34,37)	[73,69;102,58]	[24; 220]	0,0013 **
CR, кПа/л	1,08 (0,44)	[0,89;1,27]	[0,28;2,2]	0,51 (0,23)	[0,42;0,60]	[0,13;1,29]	$3 \cdot 10^{-6}$ ***

Показатели, наилучшим образом разбивающие выборку на разные классы, располагаются выше в дереве. Таким образом, наиболее информативный показатель с точки зрения разбиения обучающей выборки на подклассы, находится в вершине дерева.

Рассмотрим усеченное дерево для нашей задачи (рис. 1). Усеченным дерево называется потому, что каждое новое ветвление производится только при наличии достаточного количества объектов разных классов. В противном случае текущий узел становится листом и фиксацией небольшого процента ошибок, как например, самый правый узел в дереве на рис. 1.



Р и с . 1 . Усеченное дерево решений

Согласно рис. 1, наиболее информативным показателем для классификации является ООЛ, что согласуется с наименьшим Р-значением среди всех показателей в табл. 1. Однако, вторым по информативности показателем является ЖЕЛ, которая не разделяет две изучаемые патологии легких для всей выборки, но является наиболее информативным показателем для пациентов с ООЛ от 90.05% до 98%. Толщина прямоугольника в листьях дерева указывает на количество пациентов из обучающей выборки.

Т а б л и ц а 2

Предсказательная и распознавательная способность усеченного дерева решений

		Установленный диагноз		Предсказательная способность
		ИФА	ГХЛ	
Дерево решений	ИФА	24	3	88,89%
	ГХЛ	4	34	89,47%
Распознавательная способность		85,71%	91,89%	

Эффективность построенного классификатора можно измерить методом кросс-валидации. Основная идея этого метода заключается в том, что из обучающей выборки удаляется часть, например одна десятая всех объектов, строится классификатор, и затем на этой части проверяется правильно ли он классифицирует объекты. Этот процесс повторяется для каждой из частей. В результате оценивается количество правильно и неправильно классифицированных объектов при помощи

таблицы сопряженности (табл. 2). Если бы рассматривались два класса: «больной» и «здоровый», то можно было бы говорить о чувствительности и специфичности дерева решений как метода диагностики.

Согласно табл. 2, усеченное дерево верно классифицирует 24 пациента с ИФА и 34 пациента с ГХЛ, давая общую точность классификации 89,05%. Дерево решений верно «распознает» 24 из 28 или 85,71% пациентов с ИФА, и 34 из 37, или 91,98% пациентов с ГХЛ. В то же время, дерево решений верно «предсказывает» ИФА в 24 случаях из 27 или 88,89% и верно «предсказывает» ГХЛ в 34 случаях из 38 или 89,47%.

Алгоритм построения классификатора опирается лишь на данные, предоставляемые исследователем, поэтому не следует всегда ожидать общеизвестных эффектов и закономерностей — они могут не присутствовать в обучающей выборке. Напротив, некоторые математически обнаруженные закономерности трудно объяснить с точки зрения патофизиологии. В частности, в построенном усеченном дереве решений снижение ООЛ (менее 90,5% должной величины) является дифференциальным диагностическим критерием. Сомнительно, что ООЛ может рассматриваться как самостоятельный критерий дифференциальной диагностики ИФА и ГХЛ без учета других показателей механики дыхания.

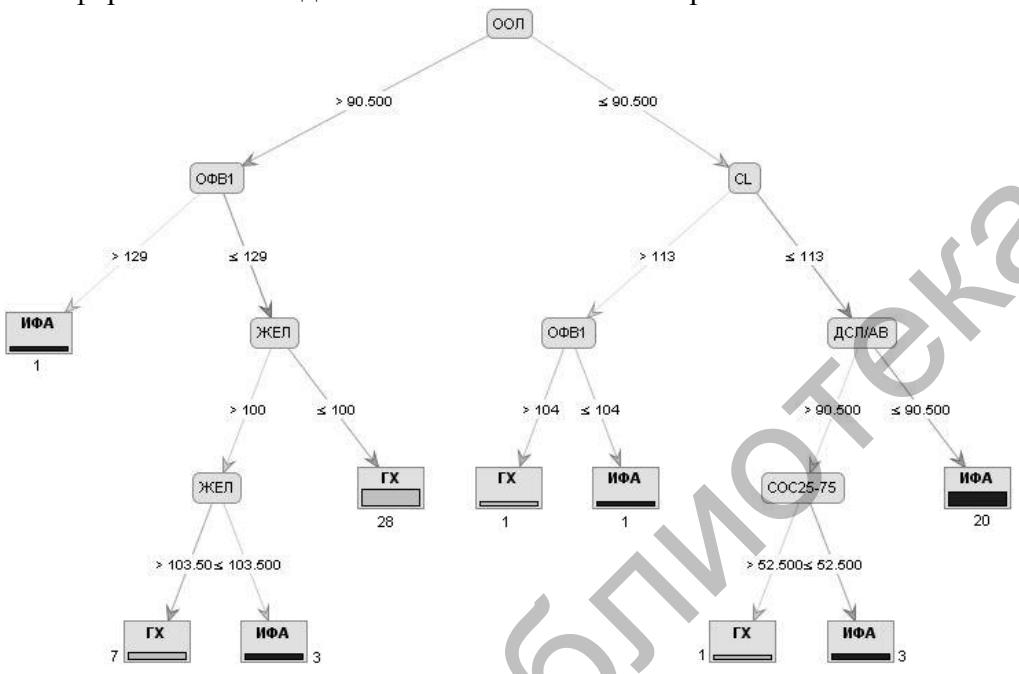
Рассмотрим теперь неусеченное дерево, представленное на рис. 2.

Главная задача алгоритма построения была найти такие параметры, которые позволят отличить ИФА от ГХЛ, пользуясь теми данными о пациентах, которые были в выборке. Правила, которые генерируются для поиска этих отличий, не обязательно будут информативными для врача-клинициста, не обязательно будут согласовываться с теоретическими представлениями и могут быть специфичными именно для данного набора пациентов. Тем не менее, некоторые правила могут представлять интерес для исследователя.

В полученном неусеченном дереве прежде всего, следует обратить внимание на правила, показывающие закономерности для явно сниженного или повышенного значения показателя. Прежде всего, это правило «если ООЛ <= 90,5 и ДСЛ/АВ <=90,5 то пациент попадает в группу ИФА». Заметим, что для дифференциальной диагностики ИФА и ГХЛ легких необходимо комплексное применение легочных функциональных тестов, включающее в себя помимо спирометрии бодиплетизмографию, определение диффузионной способности и растяжимости легких. По рисунку мы видим, что в соответствующий лист дерева решений попало наибольшее количество пациентов с ИФА.

Правила для пациентов, классифицируемых как ГХЛ, напротив, не представляют интереса, поскольку ветвление не связано со

снижением или увеличением показателей. В том числе лист, соответствующий наибольшему количеству пациентов ГХЛ, задается правилом «если ООЛ >= 90,5 и $ОФВ_1$ <=129 и ЖЕЛ < 100 то пациент попадает в группу ГХЛ». Все условия данного правила неинформативны – им соответствует как норма, так и снижение или повышение показателя. В данном случае это правило эффективно для поиска отличия пациентов ГХЛ и ИФА на нашей выборке, но неинформативно с медико-биологической точки зрения.



Р и с . 2 . Неусеченное дерево решений

В дереве также могут отобразиться неочевидные закономерности, которые достаточно сложно обнаружить на больших выборках. Например, одна из ветвей дерева говорит о том, что при сниженном ООЛ, не сниженном ДСЛ/АВ, но значительно сниженном COC_{25-75} пациент попадает в группу ИФА. Другая ветвь показывает, что есть примеры больных ИФА с повышенными ООЛ и $ОФВ_1$. И то и другое правило указывают на малое количество пациентов, поэтому, возможно, это лишь исключительные случаи. Но, тем не менее, они могут быть интересны для дальнейшего исследования.

Неусеченное дерево обладает уже не столь высокой точностью классификации – 78,33% (табл. 3).

В частности, класс ИФА, который дал нам наиболее интересные правила в дереве, был верно распознан лишь в 19 случаях из 28.

В целом, деревья решений позволяют выявить более сложные закономерности, чем стандартные статистические алгоритмы поиска

корреляции и различий между группами пациентов. При этом статистически значимые отличия между группами находят свое отражение в правилах деревьев решений.

Таблица 3

Предсказательная и распознавательная способность неусеченного дерева решений

		Установленный диагноз		Предсказательная способность
		ИФА	ГХЛ	
Дерево решений	ИФА	19	5	79,17%
	ГХЛ	9	32	78,05%
Распознавательная способность		67,86%	86,49%	

Не все закономерности, выявленные деревьями решений, представляют практический интерес. Появление некоторых неинформативных правил обусловлено, во-первых, постановкой задачи классификации с недостаточным количеством и разнообразием классов с медико-биологической точки зрения, и, во-вторых, недостаточно репрезентативной выборкой пациентов.

Тем не менее, закономерности в виде правил, предоставляемые деревом решений, могут использоваться в научных исследованиях как этап при построении алгоритмов диагностики, а также для поиска нестандартных медицинских случаев и их дальнейшего анализа.

Список литературы

1. *Каменева М.Ю.* Исследование функции внешнего дыхания // Интерстициальные заболевания легких / под ред. М.М. Ильковича, А.Н.Кокосова. СПб.: Нордмед-издат, 2005. С.50–59.
2. *Каменева М.Ю.* Стратегия применения легочных функциональных тестов в работе врача общей практики // Рос. семейные врач. 2012. Т. 6, № 2. С. 4–8.
3. *Клемент Р.Ф., Лаврушин А.А., Котегов Ю.М.* Инструкция по применению формул и таблиц должных величин основных спирографических показателей. Л., 1986. 23 с.
4. European Community for Steel and Coal: standardized lung function testing // Eur. Respir. J. 1993. Vol. 6. S. 16. P. 5–40.
5. Lung function testing / ed. By R.Gosselink, H. Stam // Eur. Respir. 2005. Vol. 10. 206 p.
6. *Quinlan J.R.* Induction of decision trees // Machine Learning 1. Kluwer Academic Publishers, 1986. P. 81–106.

APPLYING DECISION TREES TO INTERPRETATION OF RESPIRATORY MECHANICS VIOLATION AND PULMONARY GAS EXCHANGE

A.V. Tishkov[1], M.Yu. Kameneva[2], A.A. Gladskoy[1], A.I. Gunchenko[1], V.I. Trofimov[2]

[1]Saint-Petersburg Institute for Informatics and Automation of RAS
[2]Pavlov State Medical University, Saint-Petersburg

A method of pattern identifying in spirometry, bodyplethysmography, diffusion capacity and lung compliance data for patients with idiopathic pulmonary fibrosis and lung histiocytosis X is presented. It is show that complex lung function testing is necessary to diagnose idiopathic pulmonary fibrosis and lung histiocytosis X.

Keywords: *idiopathic pulmonary fibrosis, lung histiocytosis X, lung mechanics, gas exchange, data mining, decision trees.*

Об авторах:

ТИШКОВ Артем Валерьевич–кандидат физико-математических наук, старший научный сотрудник, ФГБУ Санкт-Петербургскогий институт информатики и автоматизации РАН, 199178, Санкт-Петербург, 14-я линия ВО, д. 39.

КАМЕНЕВА Марина Юрьевна–кандидат медицинских наук, руководитель лаборатории клинической физиологии дыхания НИИ Пульмонологии, ГБОУ ВПО «Санкт-Петербургский государственный медицинский университет им. И.П. Павлова» Министерства здравоохранения РФ, 197089, Санкт-Петербург, ул. Рентгена, д. 12.

ГЛАДСКОЙ Александр Александрович–аспирант ФГБУ Санкт-Петербургскогий институт информатики и автоматизации РАН, 199178, Санкт-Петербург, 14-я линия ВО, д. 39.

ГУНЧЕНКО Анна Игоревна–аспирант ФГБУ Санкт-Петербургскогий институт информатики и автоматизации РАН, 199178, Санкт-Петербург, 14-я линия ВО, д. 39.

ТРОФИМОВ Василий Иванович–доктор медицинских наук, профессор, заведующий кафедрой госпитальной терапии, ГБОУ ВПО «Санкт-Петербургский государственный медицинский университет им. И.П. Павлова» Министерства здравоохранения РФ, 197089, Санкт-Петербург, ул. Рентгена, д. 12.