УДК 004.891.3

КОМПОЗИЦИЯ ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ РАСПОЗНАВАНИЯ СТЕПЕНИ ТЯЖЕСТИ ХРОНИЧЕСКОЙ ОБСТРУКТИВНОЙ БОЛЕЗНИ ЛЕГКИХ

Н. И. Омирова[а], *преподаватель*
М. Н. Палей[б], *заочный аспирант*
Е. В. Евсюкова[б], *доктор мед. наук, профессор*
А. В. Тишков[в], *канд. физ.-мат. наук, доцент*
[а]*Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова, Санкт-Петербург, РФ*
[б]*Санкт-Петербургский государственный университет, Санкт-Петербург, РФ*
[в]*Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, РФ*

Цель: *наиболее важным вариантом диагностики степени тяжести одного из самых распространенных бронхолегочных заболеваний — хронической обструктивной болезни легких — является спирометрия. Однако она доступна не во всех лечебно-профилактических учреждениях Российской Федерации. Целью работы является построение алгоритма диагностики хронической обструктивной болезни легких без учета спирометрии.* ***Методы:*** *в качестве математического аппарата диагностики были выбраны деревья решений, на основе которых создан коллективный классификатор; в нем реализована двухуровневая схема: предварительный диагноз по первичному дереву решений уточняется на втором этапе другим деревом решений с более узкой компетенцией.* ***Результаты:*** *низкая точность классификатора может быть повышена, если матрица результатов кросс-валидации имеет блочно-диагональную структуру и классификаторы, построенные для каждого блока, имеют более высокую точность, чем исходный классификатор. Для повышения точности классификатора с результатами кросс-валидации менее 55 % предложена и опробована схема двухуровнего классификатора. На первом этапе строится первичный классификатор, предсказания которого уточняются классификаторами, построенными для диагональных блоков исходной матрицы. Предлагаемое решение позволяет улучшить точность диагностики степени тяжести хронической обструктивной болезни легких с 52,5 до 65 %.* ***Практическая значимость:*** *дифференциальная диагностика степени тяжести хронической обструктивной болезни легких может быть проведена с удовлетворительной точностью в лечебно-профилактических учреждениях, не обладающих спирометрическим оборудованием. Предлагаемый способ улучшения точности классификатора может быть применен и в других классификаторах диагностики, если удается построить набор решателей, более компетентных в узких областях, чем первичные.*

Ключевые слова *— деревья решений, коллективные классификаторы, медицинская диагностика, хроническая обструктивная болезнь легких.*

Введение

Задача диагностики с математической точки зрения представляется задачей классификации [1]. Классифицируемыми объектами являются пациенты, атрибутами объектов — антропометрические, социально-паспортные, клинико-лабораторные и другие показатели, а классами — диагнозы. Задача диагностики может заключаться в определении отсутствия или наличия заболевания, а также в определении вида заболевания или степени тяжести. В настоящей работе определяется степень тяжести хронической обструктивной болезни легких (ХОБЛ).

Рассматриваемая выборка пациентов, страдающих ХОБЛ, содержит как числовые, так и номинальные данные. Из наиболее широко известных алгоритмов классификации деревья решений [2] являются основным классификатором, работающим с номинальными данными. Когда точность классификатора оказывается ниже приемлемого уровня согласно кросс-валидации [3], ее можно повысить с помощью построения коллективов классификаторов [4]. Обычно в этом случае исходный классификатор заменяют на смесь других и вводят решающее правило получения окончательного результата на основе результатов каждого из членов коллектива. В данной работе предлагается ответ первичного классификатора в качестве исходного результата, который подлежит уточнению коллективом классификаторов.

Описание выборки и диагностическая задача

Хроническая обструктивная болезнь легких представляет серьезную проблему в современной медицине, поскольку распространенность и летальность от этого заболевания постоянно увеличиваются [5, 6]. Диагностика степени тяжести ХОБЛ в настоящее время проводится на основании результатов исследования функции внешнего дыхания (ФВД). Однако в лечебно-профилактических учреждениях не всегда имеется необходимое оборудование для ФВД. Данные о пациентах в настоящей работе не включали ФВД.

Выборка пациентов с ХОБЛ состояла из 80 пациентов четырех степеней тяжести: первой — 19 пациентов, второй — 23, третьей — 19 и четвертой — 19 пациентов. Здоровых пациентов в выборке не было, поэтому стояла задача дифференциальной диагностики степени ХОБЛ. Выборка содержала следующие клинические и лабораторные показатели: индекс курения; признаки гиперреактивности бронхов; кашель сухой; кашель продуктивный; частота сердечных сокращений; частота дыхания; подвижность легочного края при объективном осмотре пациента; степень одышки по шкале MRC; насыщение крови кислородом; индекс массы тела; количество систем органов, со стороны которых имеется патология; показатель коморбидности (общее количество заболеваний у пациента); кумулятивный рейтинговый показатель заболеваний у гериатрических пациентов (шкала Миллера); тест с 6-минутной ходьбой; показатели качества жизни по опроснику SF-36: физический компонент здоровья PH (включающий шкалы: физическое функционирование — PF, ролевое физическое функционирование — RP, интенсивность боли — BP, общее состояние здоровья — GH) и психологический компонент здоровья MH (включающий шкалы: жизненная активность — VT, социальное функционирование — SF, ролевое эмоциональное функционирование — RE, психическое здоровье — MH); лабораторные показатели: содержание в крови гемоглобина и эритроцитов, кальций, общий белок.

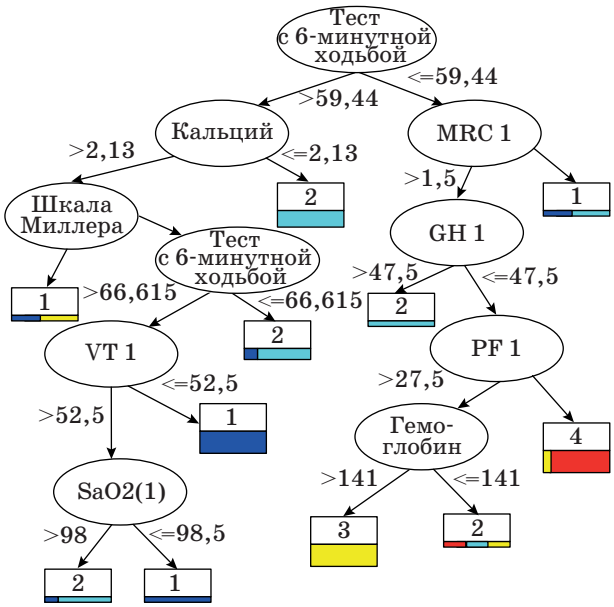
Все перечисленные показатели не относятся напрямую к симптомам бронхо-легочных заболеваний. По таким исходным данным установить тяжесть степени ХОБЛ достаточно непросто. Для достижения приемлемой точности далее предлагается схема взаимодействия классификаторов, при которой предполагаемый диагноз по первичному классификатору уточняется на втором этапе.

Двухуровневый классификатор

На первом этапе было построено дерево решений (рис. 1) для классификации пациентов по всем четырем степеням тяжести ХОБЛ.

Дерево условно делит все обучающие примеры на две группы. Правое, относительно корня, поддерево включает в основном пациентов с высокой степенью тяжести ХОБЛ: больше всего пациентов с четвертой и третьей степенью и лишь несколько пациентов со второй и первой. Левое поддерево включает обучающие примеры только первой и второй степени тяжести ХОБЛ.

В результате кросс-валидации была достигнута точность классификации (52,5±16,5) % (табл. 1). Такой уровень точности следует признать посредственным.



■ ***Рис. 1.*** Дерево решений по четырем классам

■ ***Таблица 1.*** Результат кросс-валидации по четырем классам

Предполагаемый класс и точность предсказания	Фактический класс				Точность распознавания, %
	1	2	3	4	
Класс 1	10	9	0	0	52,63
Класс 2	9	9	4	1	39,13
Класс 3	0	3	8	3	57,14
Класс 4	0	2	7	15	62,50
Точность предсказания, %	52,63	39,13	42,11	78,95	–

Ошибки классификации возникают в большей степени между классами 1 и 2, 2 и 3, 3 и 4, в меньшей степени — между 2 и 4 и не возникают между 1 и 3, 1 и 4.

Блочно-диагональный вид матрицы предсказаний порождает гипотезу о возможности построения классифиторов, анализирующих пары соседних классов: первого со вторым, второго с третьим и третьего с четвертым. Эти классификаторы будут использоваться для уточнения первичной классификации по четырем классам.

Для каждого класса строится цепочка классов, «достижимых» из данного. Достижимость означает наличие соответствующих ошибок между рассматриваемыми классами при кросс-валидации.

Например, из класса 1 достижим только класс 2, для которого имеется девять ложноположительных результатов. Ложноположительные результаты для класса 1 относительно классов 3 и 4

отсутствуют. Таким образом, для класса 1 имеем короткую цепочку 1–2. Поэтому если первичный классификатор покажет класс 1, то будет использован только один уточняющий классификатор 1–2.

Из класса 2 достижимы классы 1, 3 и 4. Поэтому цепочка для класса 2 будет выглядеть следующим образом: 1–2–3–4. Соответственно, при классификации будут использованы уточняющие классификаторы 1–2, 2–3 и 3–4. Движение по цепочке возможно от класса 2 в меньшую сторону к классу 1 по результату классификатора 1–2 или в большую сторону к классу 3 по результату классификатора 2–3 и, возможно, к классу 4 по результату классификатора 3–4. Если на первом шаге классификаторы 1–2 и 2–3 показали разнонаправленное движение, необходимо остановиться на классе 2. Движение по цепочке останавливается, если очередной уточняющий классификатор подтверждает результат предыдущего. Аналогично для класса 3 строится цепочка 2–3–4, а для класса 4 — цепочка 2–3–4.

Далее необходимо убедиться в достаточной точности уточняющих классификаторов.

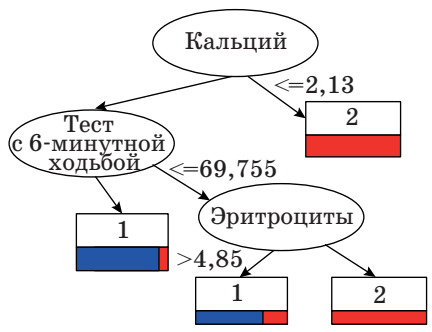
Классификатор 1–2 (рис. 2) строится на обучающей выборке, включающей два класса (первый — 1-я степень ХОБЛ и второй — 2-я степень ХОБЛ).

В качестве корневого атрибута дерева решений выступает показатель — содержание кальция в крови пациента. В результате кросс-валидации достигнута точность классификации (75±19,4) % (табл. 2).

Заметим, что суммарное количество ошибок распознавания классов 1 и 2 уменьшилось с 23 до 10 по сравнению с первичным классификатором, что составляет 31 % от общего количества пациентов класса 1 и 2.

Точность представлена в табл. 3 согласно кросс-валидации всех уточняющих классификаторов.

Поскольку распознавательная способность уточняющих классификаторов выше исходного, можно предположить, что двухуровневая система — первичный классификатор плюс уточняющий — повысит точность классификации.



■ ***Рис. 2.*** Классификатор для классов 1 и 2

■ ***Таблица 2.*** Таблица точности классификатора по кросс-валидации

Предполагаемый класс и точность предсказания	Фактический класс		Точность распознавания, %
	1	2	
Класс 1	13	4	76,47
Класс 2	6	19	76,00
Точность предсказания, %	68,42	82,61	–

■ ***Таблица 3.*** Точность уточняющих классификаторов

Уточняющий классификатор	Точность уточняющего классификатора, %	Повышение точности относительно первичного классификатора (по соответствующим классам), %
1–2	75,00±19,40	31
2–3	64,50±26,50	25
3–4	84,17±17,26	23

■ ***Таблица 4.*** Результат кросс-валидации двухуровневого классификатора

Предполагаемый класс и точность предсказания	Фактический класс				Точность распознавания, %
	1	2	3	4	
Класс 1	14	5	0	0	73,68
Класс 2	6	11	2	2	52,38
Класс 3	0	4	13	4	61,90
Класс 4	0	0	5	14	73,68
Точность предсказания, %	70,00	55,00	65,00	70,00	–

Для анализа точности двухуровневого классификатора также была проведена кросс-валидация (табл. 4).

Точность двухуровневого классификатора составила 65 %, что на 12,5 % (что соответствует десяти примерам) выше точности первичного классификатора.

Заключение

Низкая точность классификатора может быть повышена, если матрица результатов кросс-валидации имеет блочно-диагональную структуру и классификаторы, построенные для каждого блока, имеют более высокую точность, чем исходный классификатор.

Для повышения точности классификатора с результатами кросс-валидации менее 55 % предложена и опробована схема двухуровневого классификатора. На первом этапе строится первичный классификатор, предсказания которого уточняются классификаторами, построенными для диагональных блоков исходной матрицы.

При помощи двухуровневого классификатора точность распознавания степени тяжести ХОБЛ была повышена на 12,5 % и составила 65 %.

Литература

1. **Дюк В., Эмануэль В.** Информационные технологии в медико-биологических исследованиях. — СПб.: Питер, 2003. — 528 с.
2. **Quinlan J. R.** C4.5: Programs for Machine Learning. — Morgan Kaufmann Publishers, 1993. — 302 p.
3. **Воронцов К. В.** Комбинаторный подход к оценке качества обучаемых алгоритмов// Математические вопросы кибернетики. 2004. Т. 13. С. 5–36.
4. **Воронцов К. В.** Лекции по алгоритмическим композициям. http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf (дата обращения: 29.05.2014).
5. **Chapman K. R., Mannino D. M., Soriano J. B.** Epidemiology and Costs of Chronic Obstructive Pulmonary Disease//Eur. Respir J. 2006. Vol. 27. N 1. P. 188–207.
6. **Global** Initiative for Chronic Obstructive Lung Disease — GOLD. http://www.goldcopd.org/uploads/users/files/GOLD_Report_2013.pdf (дата обращения: 29.05.2014).

UDC 004.891.3

Composition of Decision Trees for Severity of Chronic Obstructive Pulmonary Desease Recognition

Omirova N. I.[a], Lecturer, nargiz.eubova@spb-gmu.ru
Paley M. N.[b], Post-Graduate Student, mnpaley@mail.ru
Evsyukova H. V.[b], Dr. Sc., Med., Professor, eevs@yandex.ru
Tishkov A. V.[c], PhD, Phys.-Math., Associate Professor, artem.tishkov@gmail.com
[a]Pavlov First Saint-Petersburg State Medical University, 6/8, L'va Tolstogo St., 197002, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg State University, 7-8, Universitetskaya Emb., 197198, Saint-Petersburg, Russian Federation
[c]Saint-Petersburg Institute for Informatics and Automation of RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation

Purpose: Chronic obstructive pulmonary disease is one of the most prevalent pulmonary diseases, and spirometry is one of the most important methods to diagnose its severity. Unfortunately, spirometry is not widely available in Russian hospitals and clinics. This paper proposes an algorithm of COPD severity diagnostics without spirometry. **Methods:** As a mathematical framework for the diagnostics, decision trees were chosen. On their base, a two-level compositional classifier was implemented. The primary decision tree provides a preliminary diagnosis refined in the second step by another more specialized decision tree. **Results:** The low accuracy of the classifier can be improved if two conditions are met: the confusion matrix has block-diagonal structure, and the classifiers built for each block have a higher accuracy than the original classifier. In order to improve the cross-validation accuracy of the classifier from less than 55%, a two-level classifier scheme is proposed and tested. First-level classifier is refined by a number of secondary classifiers built for the diagonal blocks of the original confusion matrix. The proposed solution improves the accuracy of the COPD severity diagnostics from 52,5 to 65%. **Practical relevance:** The differential diagnostics of COPD severity can be performed with satisfactory accuracy even in hospitals without spirometry equipment. The proposed method for improving the classifier accuracy can be applied in other diagnostics classifiers, if a set of solvers more competent in narrow areas than the original ones is successfully built.

Keywords — Decision Trees, Compositional Classifiers, Medical Diagnostics, COPD.

References

1. Duke V., Emanuel V. *Informatsionnye tekhnologii v mediko-biologicheskikh issledovaniiakh* [Information Technologies in Biomedical Research]. Saint-Petersburg, Piter Publ., 2003. 528 p. (In Russian).
2. Quinlan J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993. 302 p.
3. Vorontsov K. V. Combinatorial Approach to Evaluating the Quality of Trained Algorithms. *Matematicheskie voprosy kibernetiki*, 2004, vol. 13, pp. 5–36 (In Russian).
4. Vorontsov K. V. *Lektsii po algoritmicheskim kompozitsiiam* [Lectures on Algorithmic Compositions]. Available at: http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf (accessed 29 May 2014).
5. Chapman K. R., Mannino D. M., Soriano J. B. Epidemiology and Costs of Chronic Obstructive Pulmonary Disease. *Eur. Respir J.*, 2006, vol. 27, no. 1, pp. 188–207.
6. *Global Initiative for Chronic Obstructive Lung Disease — GOLD*. Available at: http://www.goldcopd.org/uploads/users/files/GOLD_Report_2013.pdf (accessed 29 May 2014).