УДК 519.24:616-006.61

ПРОГНОЗИРОВАНИЕ УСПЕШНОСТИ ЛЕЧЕНИЯ РАКА ЛЕГКОГО С ПРИМЕНЕНИЕМ АНСАМБЛЕЙ КЛАССИФИКАТОРОВ

А. Б. Корховая
Магистрант БГУИР, инженер-программист IBA IT Park

М. С. Абрамович
Заведующий НИЛ статистического анализа и моделирования Научно-исследовательского института прикладных проблем математики и информатики БГУ, кандидат физико-математических наук, доцент

Научно-исследовательский институт прикладных проблем математики и информатики БГУ, Республика Беларусь.
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь.
Иностранное предприятие «АйБиЭй АйТи Парк», Республика Беларусь.
E-mail: korhi1536@gmail.com.

А. Б. Корховая
Окончила Белорусский государственный университет информатики и радиоэлектроники. Магистрант БГУИР. Работает в IBA IT Park в должности инженера-программиста. Проводит научные исследования в области диагностики заболеваний с использованием методов машинного обучения.
М. С. Абрамович
Заведующий НИЛ статистического анализа и моделирования Научно-исследовательского института прикладных проблем математики и информатики БГУ, кандидат физико-математических наук, доцент. Область научных интересов – методы и алгоритмы машинного обучения для диагностики заболеваний.

Аннотация. Рассмотрены отличительные особенности ансамблей классификаторов: градиентного бустинга и случайного леса и условия их применения. На обучающей выборке больных раком легкого, состоящей из групп успешно излеченных больных и больных с летальным исходом, найдены значения гиперпараметров ансамблей классификаторов, обеспечивающих наибольшую точность классификации. Результаты классификации экзаменационной выборки показали высокую эффективность прогнозирования успешности лечения больных раком легкого.

Ключевые слова: рак легкого, ансамбли классификаторов, случайный лес, градиентный бустинг, гиперпараметры, точность классификации.

Введение.

В настоящее время актуальной проблемой является прогнозирование успешности лечения рака легкого на основании таких показателей как стадия рака, возраст пациента, проведения химиотерапии, лучевой терапии и ряда других. Для решения этой задачи применяются различные методы машинного обучения, в частности, деревья решений [1].

Результаты классификации с применением деревьев решений являются неустойчивыми и зависят от параметров алгоритма, сбалансированности объемов классов и объема обучающей выборки [2]. При использовании деревьев решений может также возникнуть проблема переобучения, когда число признаков сравнимо или больше числа наблюдений.

Для решения вышеуказанных проблем и повышения точности и устойчивости классификации широко применяются ансамбли деревьев решений, основанные на технологиях бустинга и бэггинга [2]: градиентный бустинг и случайный лес. В настоящей работе показано применение этих ансамблей классификаторов для прогнозирования успешности лечения рака лёгкого.

Материалы и методы.

В работе анализировались данные, предложенные экспертом в области лечения рака лёгкого. Выборка представляла собой данные о больных раком легкого Минского района, которая включают в себя 380 пациентов (232 из которых были успешно излечены и 148 с летальным исходом) и более 100 признаков.

Однако для такого объёма данных количество признаков достаточно большое и часть из них не являются информативными для построения решающих правил классификации. Поэтому экспертом были отобраны 11 признаков, которые могут потенциально быть использованы для построения ансамблей классификаторов.

Далее для этого набора признаков с целью отбора из них информативных применялся критерий хи-квадрат. В результате для построения ансамблей алгоритмов классификации было отобрано 8 информативных признаков: проводилась ли лучевая терапия (0 – не проводилась, 1 – проводилась), проводилась ли химиотерапия (0 – не проводилась, 1 – проводилась), какая операция проводилась по удалению раковой опухоли (0 – пневмонэктомия, 1 – лобэктомия/билобэктомия, 2 – резекция легкого, 3 – другие, 4 – не делали операцию), были ли жалобы у пациента (0 – отсутствовали, 1 – присутствовали), гистологическая классификация рака лёгкого (1 – плоскоклеточный, 2 – аденокарцинома, 3 – железисто-плоскоклеточный, 4 – недифференцированный, 5 – крупноклеточный, 6 – мелкоклеточный), стадия рака, диагностированная пациенту, без уточнения (1 – первая стадия, 2 – вторая стадия, 3 – третья стадия, 4 – четвертая стадия).

Для построения ансамблей классификаторов используются два основных конкурирующих подхода – бэггинг и бустинг [1, 2]. Первый из них состоит в построении множества независимых между собой моделей классификации с дальнейшим принятием решения путем голосования. Бустинг, в противоположность бэггингу, обучает каждую следующую модель с использованием данных об ошибках предыдущих моделей.

Подход бэггинга реализован в методе случайного леса [3]. Метод случайного леса обеспечивает повышение точности классификации за счет повышения независимости деревьев решений. Независимость деревьев решений достигается, во-первых, за счет того, что каждое дерево строится по выборке, получаемой из исходной обучающей выборки с помощью бутстрепа (т. е. случайной выборки с возвращением) и, во-вторых, за счет того, что при разбиении вершин при построении дерева используется только часть случайно отбираемых признаков из всего множества признаков.

Идея бустинга состоит в том, что классификаторы ансамбля строятся последовательно и на каждой итерации происходит перевзвешивание (коррекция) наблюдений, таким образом, чтобы соответствующий классификатор делал меньше ошибок на тех наблюдениях, на которых часто делали ошибки классификаторы, построенные на предыдущих итерациях алгоритма.

При применении градиентного бустинга значение параметра, регулирующего скорость обучения алгоритма, задавалось равным 0.1, так как при таком или меньшем значении параметра точность классификации увеличивается на каждом шаге алгоритма [2].

Подбор гиперпараметров, при которых была достигнута наибольшая точность классификации каждым ансамблем, проводился по сетке с использованием 3-х кратной перекрестной проверки. В таблице 1 представлены гиперпараметры и сетка значений по которой происходи их подбор.

Таблица 1. Сетка для подбора гиперпараметров

Параметр	**Значения сетки**
Количество деревьев	10, 60, 110, 160, 210
Максимальная глубина каждого дерева	2, 3, 4, 5
Минимальное количество наблюдений для разбиения внутренней вершины	2, 3, 4

Точность классификации успешно излеченных больных и больных с летальным исходом определялась путем классификации экзаменационной выборки с набором гиперпараметров, обеспечивающим наибольшую точность классификации. Для каждого ансамбля классификаторов

вычислялись диагностическая чувствительность (ДЧ) – доля правильно классифицированных вылеченных больных, диагностическая специфичность (ДС) – доля правильно классифицированных больных с летальным исходом и диагностическая эффективность (ДЭ) – доля правильно классифицированных пациентов всей выборки.

Результаты.

Объем экзаменационной выборки составлял половину всей выборки, а объем выборки наблюдений, которые использовались для подбора оптимальных гиперпарамеров ансамблей, полагался равным 80 % от объема обучающей выборки. Эта выборка формировалось с применением процедуры бутстрепа [2].

В таблице 2 для ансамблей классификации приведены значения гиперпараметров, при которых была достигнута наибольшая точность классификации обучающей выборки.

Таблица 2. Значения гиперпараметров, при которых достигнута наибольшая точность классификации

Ансамбль классификации	**Количество деревьев**	**Максимальная глубина каждого дерева**	**Минимальное количество наблюдений для разбиения внутренней вершины**
Градиентный бустинг	60	3	3
Случайный лес	110	5	3

Приведенные в таблице 2 значения гиперпараметров были использованы при классификации экзаменационной выборки и использованием случайного леса и градиентного бустинга. В таблице 3 приведены показатели эффективности для каждого ансамбля, полученные путем классификации экзаменационной выборки.

Таблица 3. Показатели точности прогнозирования с использованием методов классификации для экзаменационной выборки

Ансамбль классификации	**ДЧ**	**ДС**	**ДЭ**
Градиентный бустинг	78.0%	80.0%	78.4%
Случайный лес	86.3%	82.50%	85.5%

Как следует из таблицы 3, наибольшая точность прогнозирования успешности лечения больных раком легкого была достигнута на экзаменационной выборке с применением метода случайного леса и составила 86.3 % на экзаменационной выборке. Этот же метод оказался более эффективным при классификации всей выборки больных раком легкого (85.5 %).

Заключение.

Полученные результаты показывают, что ансамбли классификаторов показали высокую эффективность при прогнозировании успешности лечения больных раком легкого.

Список литературы

[1] Harrington, P. Machine Learning in Action / P. Harrington. New York: Manning, 2012. 382 p.

[2] Hastie, T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. H Friedman. 2nded. New York, Springer Publ., 2009. 764 p.

[3] Чистяков, С. П. Случайные леса: обзор / С. П. Чистяков // Труды Карельского научного центра РАН. 2013. №1. С. 117-136.

PREDICTION OF THE SUCCESS OF LUNG CANCER TREATMENT USING ENSEMBLES OF CLASSIFIERS

A.B. KARKHAVAYA
Graduate student of the BSUIR, software engineer
IBA IT Park

M.S. ABRAMOVICH,
Candidate of Physical and Mathematical Sciences
Head of the Research Laboratory Statistical Analysis and Modeling of Research Institute of Applied Mathematics and Informatics Problems

Research Institute for Applied Problems of Mathematics and Informatics of the Belarusian State University, Republic of Belarus
Belarusian State University of informatics and radio electronics, Republic of Belarus
IBA IT Park, Republic of Belarus
E-mail: korhi1536@gmail.com

Abstract. The distinctive features of ensembles of classifiers: gradient boosting and random forest and the conditions for their application are considered. On a training sample of patients with lung cancer, consisting of groups of cured patients and patients with a fatal outcome, the values of the hyperparameters of ensembles of classifiers that provide the highest classification accuracy were found. The results of the classification of the examination sample showed a high efficiency of predicting the success of the treatment of patients with lung cancer.

Keywords: lung cancer, ensembles of classifiers, random forest, gradient boosting, hyperparameters, classification accuracy.