

ДЕРЕВЬЯ РЕШЕНИЙ КАК МЕТОД ДИАГНОСТИКИ ЗЛОКАЧЕСТВЕННЫХ НОВООБРАЗОВАНИЙ

Хачатрян А. Г.

Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: arman_vir@mail.ru

В данной статье представлены актуальность и обоснование выбора данной тематики. Рассматривается метод классификации объектов под название «Деревья решений», а так же описаны главные преимущества выбранного метода. Предоставлены результаты основных этапов экспериментальной части исследования. И сделано заключение на основе всей проделанной работы.

ВВЕДЕНИЕ

Меланома кожи является одной из самых опасных злокачественных новообразований человека. Показатели заболеваемости увеличиваются в 2 раза среди населения Европы каждые 10-20 лет (ежегодный прирост составляет 3% - 7%) [1]. В Европе меланома занимает 17 место среди наиболее часто диагностируемых видов рака у мужчин и 8 место у женщин [2]. Со временем о меланоме сложилось мнение как о заболевании с вариабельным и неблагоприятным клиническим протеканием. В отличие от других злокачественных новообразований, меланома чаще поражает людей относительно молодого возраста, обладает высоким метастатическим потенциалом и резистентностью к лучевой и химиотерапии. Хотя только 4% от всех злокачественных новообразований кожи является меланома, на неё приходится 79% смертей этой группы. Несмотря на это, локальная меланома не является болезнью с однозначно плохим прогнозом. В случае ранней диагностики, поддающимися лечению оказываются до 95% опухолей.

I. АКТУАЛЬНОСТЬ

Лечение больных меланомой кожи является одной из самых актуальных проблем онкологии. Среди нерешенных вопросов в лечении больных меланомой кожи остаётся и вопрос о необходимости проведения профилактической регионарной хирургической операции по удалению лимфоузлов.

На протяжении долгого времени учёные пытаются определить факторы, которые влияют на протекание и точность прогнозов заболевания у больных меланомой кожи. В 2002 г утвердили новую версию классификации меланомы кожи [3]. Изменения были произведены на основании анализа прогностических факторов, изученных у 17 600 больных меланомой, и коснулись следующих параметров:

- толщина опухоли, наличие или отсутствие изъязвления;

- количество метастатически поражённых лимфоузлов и наличие микро- или макро-метастазов в лимфоузлах;
- наличие сателлитов вокруг первичной опухоли и транзиторных метастазов (было объединено в единую III стадию заболевания);
- дополнительную информацию о стадии заболевания получают при лимфатическом картировании и биопсии сторожевого лимфоузла.

II. ОПИСАНИЕ МЕТОДА

Деревья решений – метод, который позволяет устанавливать принадлежность объектов к тому или иному классу зависимой категориальной переменной, измеренной в списке наименований, в зависимости от сопутствующих значений одной или нескольких предикативных переменных [4]. Главной целью построения дерева классификации является построение решающих правил, позволяющих классифицировать объекты. Способность деревьев классификации выполнять параллельное ветвление для анализа вклада отдельных переменных дает возможность работать с предикативными переменными различных типов – как с категориальными, так и с непрерывными. Деревья классификации могут быть очень сложными. Однако графическое представление, позволяет упростить понимание результатов даже для довольно сложных деревьев. Возможность графического представления результатов и простота интерпретации во многом объясняют достаточно большую распространённость деревьев классификации в медицине, а также и в других областях. Процесс построения дерева классификации делится на 4 этапа:

1. определение критерия точности прогноза;
2. определение момента остановки ветвления;
3. выбор вариантов ветвления;
4. определение «необходимого размера» дерева.

Целью проведения анализа с помощью деревьев классификации, является получение максимально точного прогноза. Наиболее точным прогно-

зом является тот, который с наименьшей долей вероятности неправильно классифицирует наблюдения (цена). Соответственно самый лучший прогноз – это тот, который дает наименьшую цену [4].

III. ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

В исследование вошёл 716 больной первичной меланомой кожи с локализацией опухоли на коже туловища и конечностей, получавший лечение в НИИОМР за период 1965-2002 гг. В основу деления больных на группы был положен факт развития метастазов в регионарных лимфоузлах. Средний возраст больных, включенных в исследование, составил 49 лет (в пределах 10-84 года). В обеих группах преобладали женщины. В первой группе больные меланомой с метастазами (19,4%), во второй – без метастазов (80,6%). Для определения влияния факторов на развития метастазов в регионарных лимфатических узлах было отобрано 23 клинико-морфологических признака, такие как пол, возраст, кровоточивость опухоли, толщина опухоли по Breslow и т.д.

Для отнесения пациентов к одной из групп «с метастазами» или «без метастазов», использовались решающие правила, построенные с использованием деревьев решений. Метод классификации деревьев решений позволяет предсказывать принадлежность наблюдений к тому или иному классу категориального признака в зависимости от соответствующих значений одного или нескольких предикторных признаков.

Для построения дерева решений использовался алгоритм C&RT.

В алгоритме C&RT каждая вершина дерева представляет собой правило классификации связанное с одним признаком, который наилучшим образом разделяет обучающую выборку на два класса. Для выбора наилучшего из всех возможных вариантов ветвления (т.е. выбора признаков, по которым строилось дерево решений) использовался критерий Джини. По критерию Джини в качестве признака ветвления выбирается такой признак, который минимизирует значение критерия в данной вершине.

Таблица 1 – Важность признаков для классификации

Название	Ранг	Важность
Длительность заболевания до начала лечения	100	1.000000
Площадь	96	0.958123
Толщина опухоли по Breslow, мм	81	0.805809
Уровень инвазии по Clark	44	0.440096
Изъязвление	32	0.323073
Код роста опухоли	20	0.204083
Врастание опухоли в сосуды	19	0.191236

Выбор оптимального размера дерева решений определялся с помощью кросс-проверки. Этот вид проверки рекомендуется использовать в том случае, когда нет отдельной экзаменационной выборки, а обучающая выборка является недостаточно большой, чтобы из неё можно было выделить экзаменационную. В нашем исследовании применялась 3-кратная кросс-проверка. При такой процедуре из обучающей выборки случайным образом формировались 3 подвыборки такого же объёма, как и обучающая выборка (использовался подход построения выборки с возвращением). Дерево решений нужного размера строилось 3 раза, причем каждый раз по очереди одна из подвыборок не использовалась при его построении, а рассматривалась как экзаменационная и для неё вычислялась доля ошибочной классификации. Доли ошибочной классификации по трём экзаменационным выборкам усреднялись и, таким образом, определялась доля ошибочной классификации 3-кратной кросс-проверки.

Таблица 2 – Результаты эксперимента

Класс(всего)	Правильно	Неправильно
С метастазами(139)	111	28
Без метастазов(577)	439	138

Таким образом из таблицы 2 можно увидеть что метод «Дереья решений» правильно классифицировал 80%(диагностическая чувствительность) больных с метастазами, 76%(диагностическая специфичность) больных без метастазов и 77%(диагностическая эффективность) от общего количества больных.

ЗАКЛЮЧЕНИЕ

В данной статье рассматривается метод «Дереья решений», для классификации больных меланомой кожи на две группы:

- больные меланомой с метастазами;
- больные меланомой без метастазов.

Показана важность выбора информативных признаков и размера дерева при построении решающего правила для классификации.

СПИСОК ЛИТЕРАТУРЫ

1. Osterlind A. Epidemiology on malignant melanoma in Europe. // Acta. Oncol. – 1992. – Vol. 31, № 8. – P. 903–908.
2. CLOBOCAN 2000: cancer incidence, mortality and prevalence worldwide, version 1.0. Ferlay J., Bray F., Pisani P., Parkin DM. // IARC cancerbase – 2001. – №.5.
3. Prognostic factors analysis of 17,600 melanoma patients. Validation of the AJCC melanoma staging system. Balch C.M., Soong S.J., Gershenwald J.E., et al. // J. Clin. Oncol. – 2001. – Vol. 19. – P. 3622-3634.
4. Деревья классификации как метод определения тяжести больных, страдающих хронической сердечной недостаточностью. А.А. Халафян, Л.Н. Елисева, С.Г. Сафонова, Савенко Д. В. – 2006. – № 11