https://doi.org/10.26160/2474-5901-2019-17-161-166

ФОРМИРОВАНИЕ ДИАПАЗОНОВ ПЕРЕМЕННЫХ ЭКСПЕРТНОЙ СИСТЕМЫ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВА ПРИНЯТИЯ РЕШЕНИЙ

Серобабов А.С.

Ключевые слова: классификация данных, дерево решений, энтропия Шеннона, ранняя диагностика.

Аннотация. В статье рассматривается задача классификации стадии заболевания в экспертной системе по входным лабораторным параметрам пациента. Описаны предварительные этапы по подготовке данных к построению дерева принятия решений. Реализован алгоритм построения бинарного дерева, где оценка качества модели, произведена с помощью информационной меры Шеннона. В результате получено дерево правил классификации стадии заболевания по входным параметрам системы.

FORMATION OF RANGES OF VARIABLES OF EXPERT SYSTEM USING DECISION TREE

Serobabov A.S.

Keywords: data classification, decision trees, Shannon entropy, early diagnosis.

Abstract. The article deals with the problem of classification of the disease stage in the expert system according to the input laboratory parameters of the patient. The preliminary stages of data preparation for decision tree construction are described. An algorithm for constructing a binary tree is implemented, where the assessment of the quality of the model is made with the help of the Shannon information measure. As a result, a tree of rules for classification of the disease stage by the input parameters of the system is obtained.

За последние десятилетия сделаны значительные успехи в сфере информационных технологий. Быстрое развитие вычислительных способностей компьютера и разработка новых алгоритмов анализа данных сегодня позволяют решать задачи, которые раньше являлись не выполнимыми на базе ЭВМ. К таким задачам можно отнести раннее диагностирование заболевания, прогнозирование развития болезни и распознавание очагов воспаления – все они и в настоящее время остаются актуальными для исследовательской деятельности.

Данная работа посвящена задаче формирования диапазонов переменных экспертной системы для классифицирования пациентов по стадиям заболевания печени с использованием метода дерева принятия решения. Так как процедура диагностики это выбор последовательной оценки входных параметров и общего физиологического состояния пациента, а диапазоны принадлежности количественных параметров имеют нечетко выраженные отличия, то имеет место быть множество решений классификации, среди которых необходимо выбрать оптимальный вариант. Данная статья является продолжением цикла статей [1, 2], посвященных проблеме ранней диагностики заболеваний. На сегодняшний день существует значительное число алгоритмов, реализующих деревья решений: CART (классификация и регрессионное дерево), ID5, CN2, CHAIN и др. [3]. Основной же раздел

данной статьи посвящен реализации алгоритма классификации методом CART, так как этот алгоритм является непараметрическим, поэтому нет необходимости рассчитывать различные параметры вероятностного распределения, заранее выбирать переменные, которые будут участвовать в анализе, так как сами переменные отбираются непосредственно во время проведения анализа на основании значения энтропии Шеннона. Для применения алгоритма CART нет необходимости принимать дополнительные предположения или допущения перед проведением анализа.

На основании данных, полученных из медицинских учреждений города Омск, собранных посредством диспансеризации населения, рассмотрена задача, когда необходимо на основе входных параметров экспертной системы диагностики заболевания принять решение о принадлежности пациента к одной из стадий болезни. Для этого введем следующие обозначения входных параметров исследования: L_{obr} – количество рецепторов воспринимающих лептин, L_{ler} – содержание лептина в организме пациента, D_{el} – стадия заболевания неалкогольной жировой болезни печени.

На рисунке 1 изображена схема классификации заболевания по целевой переменной (переменная, которая описывает результат (цель) процесса). На этапе 1 отбираются значимые параметры L_{obr} и L_{ler} для того, чтобы отнести пациента к определенной стадии заболевания. На этапе 2 из выборки исключаются пациенты, у которых отсутствует один из значимых параметров исследования. На этапе 3 производится трансформация данных для дальнейшей передачи их в функцию, отвечающую за построение решающего дерева.

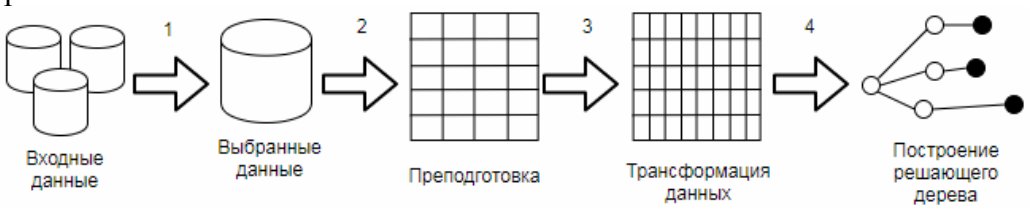

Рис. 1. Этапы классификации заболевания по целевой переменной (стадия заболевания)

Этап 4 – на вход функции построения решающего дерева поступает множество входных данных $A = U_1^m\left(\left\{L_{lep}, L_{obr}, D_{el}\right\}_i\right)$, где D_{el} – зависимая целевая переменная (стадия заболевания, которую необходимо классифицировать), i – индекс пациента, m – количество пациентов тренировочной выборки, $L_{lep}, L_{obr} \in R$. В качестве оценочной функции алгоритма CART используется метод энтропии Шеннона, базирующийся на идее прироста информации при разбиении на новых два узла дерева. Ниже представлен обобщенный алгоритм CART для построения дерева решений:

Алгоритм CART с использованием энтропии

1. Ввод значения глубины построения дерева L (расстояние от корня дерева до его листа);
2. Вызвать рекурсивную функцию поиска оптимальных предикатов разбиения $F(A, L)$ (инициализирующий шаг для выполнения рекурсивной функции поиска оптимальных предикатов разбиения на два подмножества, где предикат – выражение, использующее одну или более величину с результатом булевого типа);

Функция поиска предикатов F, принимающая на входе множество T для поиска предиката и ограничения глубины построения дерева L.

F1. Инициализировать множество T для поиска локального предиката дерева решений;

F2. Вычислить значение энтропии до разбиения S_0;

F3. **Если** $L > 0$ или S_0!=0 **выполнять**

F4. **Перебрать все признаки** $x_i \in (x, Y)$ из множества A, где x – входные параметры исследуемой выборки, Y – зависимая целевая переменная;

F5. **Перебрать элементы разбиения** Q_{jx_i} из кортежа множества T по **признаку** x_i;

F6. Сгенерировать предикат θ, чтобы разбить $T \to B_1, B_2$, где $B_1(p,t) = \{m \mid m_p < t\}$ и $B_2(p,t) = \{m \mid m_p \geq t\}$, $m \in T$ t – пороговое значение предиката;

F7. Вычислить значение энтропии для одной из подгрупп: $S_{Q_{jx_i n}} = -\sum_{p=1}^{q} \frac{k_p}{m} \cdot \log_2 \frac{k_p}{m}$, где k_p – количество элементов, попадающих под атрибут разбиения Q_{jx_i} в группу p, q – количество групп разбиения равное 2, n – номер группы разбиения;

F8. Вычислить прирост информации $IG(x_i) = S_0 - \sum_{p=1}^{q} \frac{k_p}{m} \cdot S_{Q_{jx_i n}}$, где $info(S_0)$ – информация с подмножеством S_0 до разбиения, $info(Q_{jx_i q})$ – информация, связанная с подмножеством, полученным по разбиению атрибута $Q_{jx_i q}$;

F9. Найти max IG;

F11. Найденный предикат θ является частью дерева решений (добавить в дерево решений Tree);

F12. $L = L - 1$;

F13. Вызвать рекурсивно функции $F(B_1, L)$; $F(B_2, L)$;

F14. Завершить рекурсивную функцию;

3. Построить дерево решений Tree;

Данный алгоритм выполнен на языке программирования Python с использованием интерактивной среды выполнения Project Jupyter. При построении дерева решений использованы следующие библиотеки: scikit-learn для обучения дерева принятия решений CART, pandas для предподготовки данных, graphviz для построения графа. За обучение решающего дерева отвечает класс DecisionTreeClassifier, конструктор которого принимает критерий классификации, максимальную и минимальную глубину построения дерева.

На рисунке 2 изображено дерево с глубиной построения 4, полученное в результате выполнения обучения на входных данных множества *A*. Как видно из рисунка 2 на корне дерева, где узел *L=0*, входное множество разбивается на два новых подмножества по предикату $X_1 \le 22{,}077$ Данному условию удовлетворяют 12 значений из 27, которые образуют новое множество B_1, остальные 15 значений образуют второе множество B_2 по остаточному признаку. Данный рекурсивный процесс разбиения заканчивается, согласно алгоритму, при достижении ограничения глубины построения дерева решений или при значении энтропии узла 0.

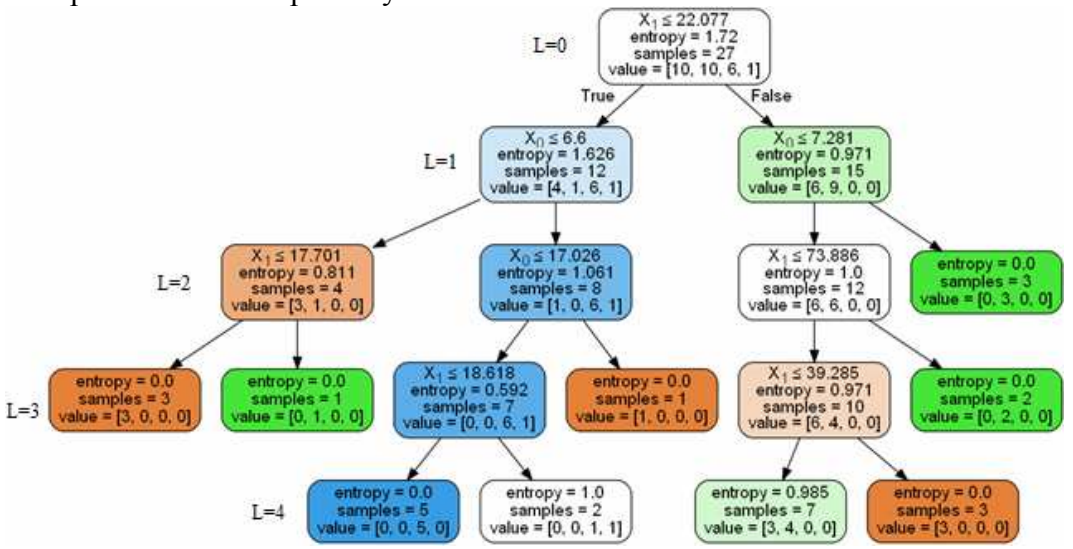

Рис. 2. Дерево принятия решений при $L=4$

На рисунке 3 изображен двумерный график разбиения на области классификации по принадлежности пациента к одному из значений D_{el} на основе обученного дерева принятий решений. Как видно из рисунка 3 решающее дерево включило в 1 стадию правила $0 > L_{lep} > 2$ && $7 > L_{lep} > 22{,}077$, куда попадает единственное значение из 1 стадии, а ближайшими соседями являются случаи, относящиеся к 3 стадии. Здесь наблюдается переобучение, что может соответствовать выбросу.

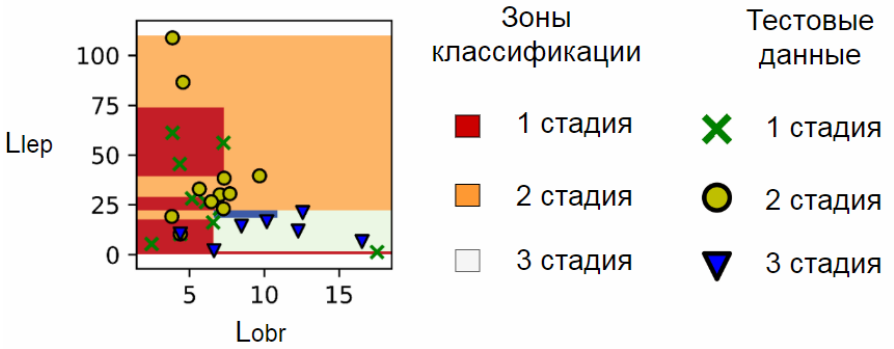

Рис. 3. Результаты применения правил при L = 4

Рассмотрен новый подход к классификации заболевания печени, при котором сначала происходит обучение модели на обучающих данных кортежа $\{L_{lep}, L_{obr}, D_{el}\}$, где целевой переменной выступает параметр D_{el} Полученное дерево принятия решений описывает структуру принимаемых решений, которыми необходимо руководствоваться, чтобы отнести пациента к одному из значений категориального параметра D_{el} На основе дерева принятия решений можно наглядно продемонстрировать эксперту в области прикладной биомедицины, какими правилами необходимо руководствоваться в ходе постановки диагноза, а также сравнить с имеющимися у эксперта заключениями. При тестировании модели получены следующие результаты: правильность классификации на тестовых данных при L=3-63%, при L=4-67%. Данные числовые значения результатов тестирования не высокой точности, поэтому необходимы дополнительные исследования в данной области. Полученные же правила разбиения могут быть применены для большего понимания прикладной области, а также для корректирования экспертных правил с учетом опытных оценок и заметок эксперта.

Список литературы

1. Серобабов А.С. Проверка входных параметров экспертной системы на соответствие нормальному закону распределения // Проблемы и перспективы студенческой науки 2019. №2(6). С. 3-6.
2. Серобабов А.С., Чебаненко Е.В., Денисова Л.А., Кролевец Т.С. Разработка экспертной системы ранней диагностики заболеваний: программные средства первичной обработки и выявление зависимостей. // Омский научный вестник. 2018. № 4 (160). С. 179-184.
3. Ларичев О.И. Качественные методы принятия решений. Вербальный анализ решений / О.И. Ларичев, Е.М. Мошкович. – М.: Наука, 1996. – 208с.

References

1. Serobabov A.S. Checking the input parameters of the expert system for compliance with the normal distribution law // Problems and prospects of student science 2019. №2(6). P. 3-6.

2. Serobabov A.S., Chebanenko E.V., Denisova L.A., Krolevets T.S. Development of an expert system for early diagnosis of diseases: software for primary treatment and detection of dependencies. // Omsk scientific Bulletin. 2018. No. 4 (160). P. 179-184.
3. Larichev O.I. Qualitative methods of decision-making. Verbal analysis of decisions / O.I. Larichev, E.M. Moshkovich. – M.: Science, 1996. – 208p.

Серобабов Александр Сергеевич – аспирант, Омский государственный технический университет, г.Омск, Россия, aserobabow95@mail.ru	**Serobabov Aleksandr Sergeyitch** – postgraduate student, Omsk state technical university, Omsk, Russia, aserobabow95@mail.ru

Received 23.12.2019