УДК 004(075.32)

Стрюков Руслан Константинович
аспирант
Воронежский Государственный Университет
Stryukov R.K.
postgraduate
Voronezh State University

РАЗРАБОТКА ДИАГНОСТИЧЕСКОГО ДЕРЕВА НА ОСНОВЕ КРИТЕРИЯ ПРИРОСТА ИНФОРМАЦИИ

DEVELOPMENT OF A DIAGNOSTIC TREE BASED ON INFORMATION GAIN CRITERION

Аннотация. *Деревья принятия решений являются гибким инструментом для решения задач диагностики. Целью данной статьи является разработка алгоритма для построения диагностического дерева.*

Ключевые слова: *дерево принятие решений, диагностическое дерево, критерий прироста информации, критерий Gini, критерий GainRatio, энтропия.*

Summary. *Decision Trees are a flexible tool for solving diagnostic tasks. The purpose of this article is to develop a diagnostic algorithm for constructing the tree.*

Keywords: *decision tree, diagnostic tree information gain criterion, the Gini criterion ratio GainRatio criterion, entropy.*

Введение

Под медицинской диагностикой будем понимать процесс установления *диагноза*, то есть заключения о сущности болезни и состоянии пациента, выраженное в принятой медицинской терминологии [1].

В качестве метода диагностики будем рассматривать дерево принятия решений или диагностическое дерево [2].

Разрабатываемый алгоритм будет решать основную проблему построения дерева: какие атрибуты необходимо выбирать для того, чтобы построенное дерево было наиболее оптимальным.

Алгоритм построения диагностического дерева

Предположим, что показатели $S_1 ... S_n$ характеризуют состояние пациента в группе заболеваний $K_1,...,K_m$.

Для каждого показателя определена шкала $SH_i (i = \overline{1,n})$ в виде конечного множества возможных значений. Мощность данной шкалы обозначим $|SH_i| = t_i$, $(i = \overline{1,n})$.

Под экземпляром будем понимать вектор i-ая компонента, которого соответствует показателю S_i. Экземпляр — вектор возможных значений показателя, характеризующий состояние пациента.

Множество, состоящее из экземпляров, каждый из которых относится к некоторому заболеванию (классу), назовем обучающим множеством LearnSet.

Алгоритм построения диагностического дерева

S1. Зафиксировать класс K_i и отобрать экземпляры $(x_1^j,...,x_n^j)$ из LS, относящиеся к данному классу $x_i^j \in SH_i$ $(i = \overline{1,h})$.

S2. Последовательно рассмотрим показатели S_j $(j = \overline{1,n})$ по каждому определить частоту появления значений из шкалы SH_i.

Пусть $sh_r^i \in SH_i$ — одно из r значений, которое встречается $|Sh_r^i|$ раз, тогда

$$p_r^i = \frac{|Sh_r^i|}{|K_i|},$$

где $|K_i|$ — количество экземпляров в данном классе K_i.

S3. Определить количество характеристик показателей для фиксированного класса K_i по каждому показателю S_j [3].

Определить энтропию по формуле

$$H(S_j \mid K_i) = -\sum_{r=1}^{|SH_j|} p_r^j \ln p_r^j .$$

S4. Процедура выбора.

Сформировать множество потенциальных корней S_j^*, включив в него показатели, выбранные на основе следующих критериев [4]:

1) Выбрать те вершины, для которых энтропия минимальна;

2) Выбираются те вершину, у которых критерий прироста информации максимален

$$Gain(A \mid SH_i) = H(A \mid K_i) - \sum_{i=1}^{q} H(A_k, S),$$

где A_k — множество элементов из A, на которых признак P имеет значение S.

3) Критерий GainRatio

$$GainRatio(A, SH_i) = \frac{Gain(A, SH_i)}{SplitInfo(A, SH_i)},$$

где

$$SplitInfo(A, SH_i) = -\sum_{i=1}^{q} \frac{|A_{sh_i}|}{|A|} \log_2 \frac{|A_{sh_i}|}{|A|}.$$

Выбираются вершины для которых GainRatio максимальный.

4) Критерий Гини

$$Gini(A, S) = 1 - \sum_{i=1}^{s} \frac{|A_i|}{A}$$

$$Gini(A, SH_i, S) = Gini(A, S) - \sum_{j=1}^{q} \frac{|A_j|}{|A|} Gini(A_j, S).$$

Тогда в качестве вершины выбирается та, у которой критерий Gini максимальный.

S5. Повторяя шаги S1–S4, строится граф в виде разложения по уровням, причем каждому уровню соответствует показатель S_i, всего уровней n. Уровни располагаются в соответствии с ранжированием показателей.

Из каждой вершины — показателя S_i, выходит столько дуг, сколько возможных значений содержит шкала SH_i.

Каждой дуге — значению Sh_r^i приписывается относительная частота появления этого значения в редуцированных экземплярах обучающего множества.

Под редуцированным экземпляром обучающего множества подразумевается экземпляр, в котором оставлены компоненты, соответствующие пути из корня в данную вершину диагностического дерева.

В результате выполнения данного шага будет построено взвешенное ориентированное дерево в виде иерархии. На нижнем уровне располагаются вершины, которые соответствуют классам заболеваний K_i.

Пример построения диагностического дерева

Пусть имеются следующие симптомы со значениями:

Головокружение: есть/нет;

Среднее содержание гемоглобина в эритроците: Сниженный (С), Нормальный (Н), Повышенный (П);

Средний диаметр эритроцитов: Сниженный (С), Нормальный (Н), Повышенный (П);

Диагнозы: Железодефицитная анемия (ЖА), Гемолитическая анемия (ГА), В12 дефицитная анемия (В12).

Имеющиеся данные представлены в таблице 1.

Таблица 1

Исходное обучающее множество

Головокружение	Среднее содержание гемоглобина в эритроците	Средний диаметр эритроцитов	Диагноз
Есть	П	П	ГА
Нет	П	П	В12
Нет	Н	Н	ГА
Нет	С	С	ЖА
Есть	П	Н	ГА
Есть	С	С	ЖА
Есть	Н	П	ГА
Есть	П	С	ЖА

Рассчитаем прирост информации

$$H(A,Диагноз) = -\sum_{i=1}^{s} \frac{m_i}{n} \ln \frac{m_i}{n} = -\frac{4}{8}\ln\frac{4}{8} - \frac{1}{8}\ln\frac{1}{8} - \frac{3}{8}\ln\frac{3}{8} = 0{,}9743$$

$$Gain(A,головокружение) = H(A,диагноз) -$$

$$- \frac{5}{8} H(головокружение\ (есть), диагноз) -$$

$$- \frac{3}{8} (головокружение\ (нет), диагноз) =$$

$$= 0{,}9743 - \frac{5}{8}\left(-\frac{3}{5}\ln\frac{3}{5} - \frac{2}{5}\ln\frac{2}{5}\right) - \frac{3}{8}\left(-\frac{1}{3}\ln\frac{1}{3} - \frac{1}{3}\ln\frac{1}{3} - \frac{1}{3}\ln\frac{1}{3}\right) = 0{,}1417$$

$$Gain(A,\ Среднее\ содержание\ гемоглобина\ в\ эритроците) =$$

$$= H(A,диагноз) - \frac{2}{8} H(ССГвЭ\ (Снижен), диагноз) -$$

$$- \frac{2}{8} (ССГвЭ\ (Нормальный), диагноз) -$$

$$- \frac{4}{8} (ССГвЭ\ (Повышенный), диагноз) =$$

$$= 0{,}9743 - \frac{2}{8}\left(-\frac{2}{2}\ln\frac{2}{2}\right) - \frac{2}{8}\left(-\frac{2}{2}\ln\frac{2}{2}\right) - \frac{4}{8}\left(-\frac{2}{4}\ln\frac{2}{4} - \frac{1}{4}\ln\frac{1}{4} - \frac{1}{4}\ln\frac{1}{4}\right) = 0{,}4544$$

$$Gain(A,\ Средний\ диаметр\ эритроцитов) =$$

$$= H(A,диагноз) - \frac{3}{8} H(СДЭ\ (Снижен), диагноз) -$$

$$- \frac{2}{8} (СДЭ\ (Нормальный), диагноз) -$$

$$- \frac{3}{8} (СДЭ\ (Повышенный), диагноз) =$$

$$= 0{,}9743 - \frac{3}{8}\left(-\frac{3}{3}\ln\frac{3}{3}\right) - \frac{2}{8}\left(-\frac{2}{2}\ln\frac{2}{2}\right) - \frac{3}{8}\left(-\frac{2}{3}\ln\frac{2}{3} - \frac{1}{3}\ln\frac{1}{3}\right) = 0{,}7356$$

Согласно вычислениям в качестве корня дерева необходимо выбрать атрибут с максимальным приростом информации — *средний диаметр эритроцитов* [5].

Поделим обучающее множество на подмножества.

Таблица 2

Обучающее множество после редуцирования (СДЭ = С)

Головокружение	Среднее содержание гемоглобина в эритроците	Средний диаметр эритроцитов	Диагноз
Нет	С	С	ЖА
Есть	С	С	ЖА
Есть	П	С	ЖА

Таблица 3

Обучающее множество после редуцирования (СДЭ = Н)

Головокружение	Среднее содержание гемоглобина в эритроците	Средний диаметр эритроцитов	Диагноз
Нет	Н	Н	ГА
Есть	П	Н	ГА

Как видно из таблиц 2 и 3 установка диагноза тривиальна, рассмотрим случай представленный в таблице 4.

Таблица 4

Обучающее множество после редуцирования (СДЭ = П)

Головокружение	Среднее содержание гемоглобина в эритроците	Средний диаметр эритроцитов	Диагноз
Есть	П	П	ГА
Нет	П	П	В12
Есть	Н	П	ГА

Рассчитаем прирост информации.

$$H(A,Диагноз) = -\sum_{i=1}^{s}\frac{m_i}{n}\ln\frac{m_i}{n} = -\frac{2}{3}\ln\frac{2}{3} - \frac{1}{3}\ln\frac{1}{3} = 0{,}6365$$

$$Gain(A,головокружение) = H(A,диагноз) -$$

$$-\frac{2}{3}H(головокружение\ (есть), диагноз) -$$

$$-\frac{1}{3}(головокружение\ (нет), диагноз) =$$

$$= 0{,}6365 - \frac{2}{3}\left(-\frac{2}{2}\ln\frac{2}{2}\right) - \frac{1}{3}\left(-\frac{1}{1}\ln\frac{1}{1}\right) = 0{,}6365$$

$$Gain(A, ССГвЭ) = H(A,диагноз) -$$

$$-\frac{0}{3}H(ССГвЭ\ (Снижен), диагноз) -$$

$$-\frac{1}{3}(ССГвЭ\ (Нормальный), диагноз) -$$

$$-\frac{2}{3}(ССГвЭ\ (Повышенный), диагноз) =$$

$$= 0{,}6365 - \frac{1}{3}\left(-\frac{1}{1}\ln\frac{1}{1}\right) - \frac{2}{3}\left(-\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2}\right) = 0{,}1744$$

Следующим критерием в случае с повышенным диаметром эритроцитов является — *головокружение*.

В результате получим следующие подмножества

Таблица 5

Обучающее множество после редуцирования (СДЭ = П и Г=есть)

Головокружение	Среднее содержание гемоглобина в эритроците	Средний диаметр эритроцитов	Диагноз
Есть	П	П	ГА
Есть	Н	П	ГА

Таблица 6

Обучающее множество после редуцирования (СДЭ = П и Г=нет)

Головокружение	Среднее содержание гемоглобина в эритроците	Средний диаметр эритроцитов	Диагноз
Нет	П	П	В12

Полученные случаи в таблицах 5 и 6 являются тривиальными. Построим диагностическое дерево.

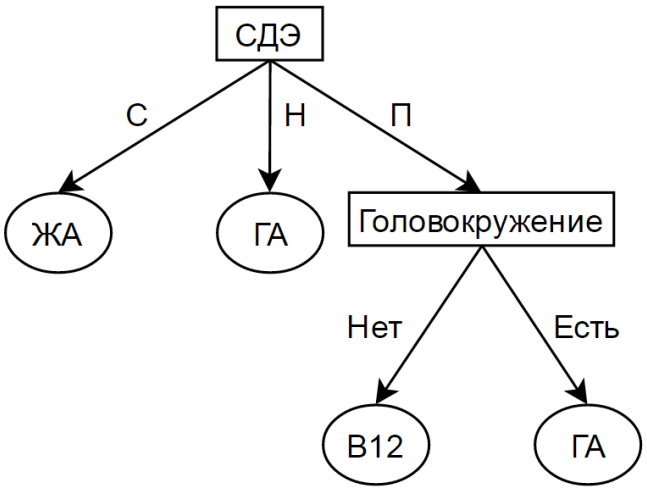

Рисунок 1. Пример диагностического дерева

Заключение

На основе разработанного алгоритма было построено диагностическое дерево, которое может быть использовано для:

– построения продукционных правил диагностики заболеваний;
– позволяет объяснить, как именно был поставлен конкретный диагноз.

Литература

1. Диагностика [электронный ресурс] URL: https://ru.wikipedia.org/wiki/Диагностика.

2. Леветин, А.В. Алгоритмы: введение в разработку и анализ.: Пер. с англ. — С.Г. Тригуб, И.В. Красикова. — Издательский дом «Вильямс», 2006. — 576 с.

3. Шеннон, К. Работы по теории информации и кибернетике.: Пер. с англ. — Р.Л. Добродушина, О.Б. Лупанова. — Москва: Издательство иностранной литературы, 1963. — 823 с.

4. Николенко, С. Деревья принятия решений [Электронный ресурс] // Персональный сайт Сергея Николенко. 2006. URL: http://logic.pdmi.ras.ru/~sergey/teaching/ml/notes-01-dectrees.pdf (дата обращения: 27.04.2015).

5. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям: учебное пособие. 2-е изд./ Н.Б. Паклин — СПб: Питер, 2013. — 444–459 с.