РАЗРАБОТКА ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ДРЕВОВИДНОЙ ИНДУКЦИИ

DATA MINING WITH USE OF ANCIENT INDUCTION

I. Kutsenko

Summary. The article considers the methodology of Data Mininig on the basis of the method of ancient induction, on the example of decision making in the insurance company. The analytical task of insurance payments is set for different groups and different life situations. The solution of the task is reduced to the implementation of the construction of the query tree and the placement of information in the root system of the tree.

Keywords: Data base, data mining, prognosis in insurance systems, the methodology of ancient induction, the root system of the prediction tree.

Куценко Ирина Львовна
К.ф.-м.н., доцент, Российский университет дружбы народов (РУДН)
i.kutsenko@mail.ru

Аннотация. В статье рассматривается методология Data Mininig на основе метода древовидной индукции, на примере принятия решения в страховой компании. Ставится аналитическая задача страховых выплат для различных групп и различных жизненных ситуаций. Решение поставленной задачи сводится к реализации построения дерева запроса и помещения информации в корневой системе дерева.

Ключевые слова: База данных, разработка данных, прогнозирование в системе страхования, метод древовидной индукции, корневая система прогнозирующего дерева.

Главной заслугой данной технологии является скорость. Благодаря своей структуре (форме куба), она позволяет выполнять запросы намного быстрее, чем обычные базы данных.

В данной статье рассмотрим алгоритм, который поможет разобраться в этих вопросах. Этот метод сегодня широко используются в сферах финансов, кредитования и страхования.

1. Древовидная индукция

Решается задача аппроксимации булевой функции методом древовидной индукции.

Пусть имеется функция $f: B^n \to B$, где $B=\{0, 1\}$ — булево множество.

$z = \{x_1, x_2, \ldots, x_n\}$, x_j — дискретная величина, обладающая свойством q.

$$f(x) = y, y_i = \begin{cases} 0 \\ 1 \end{cases}, i=1, \ldots, m$$

Тогда мы имеем

$$f(z_1) = 1$$
$$f(z_2) = 0$$
$$\vdots$$
$$f(z_n) = 0,$$

где — z_k уникальный набор иксов.

Если у нас появляется новый набор иксов, неравный ни одному предыдущему, тогда нас будет интересовать чему будет равна функция от этого набора: $f(z_{n+1}) = ?$

Алгоритм построения дерева:

1. Выбираем очередной атрибут x_j и ставим его в корень.

2. Для всех его значений q:

1. Из всех значений $f(z_k)$ оставляем только те, у которых свойство x_j равно q

2. Далее строим дерево рекурсивно этом потомке.

Согласно Т. Коннолли, древовидная индукция (также известная как деревом принятия решений) — это метод, используемый для прогнозных моделей [1].

Деревья, рассмотренные в данной статье, не отличаются от настоящих деревьев своим строением. Обычно принято читать деревья слева направо или сверху вниз, мы будем пользоваться вторым правилом.

Пусть имеется множество A состоящее из n элементов, у m из которых есть определенное свойство X. Тогда энтропия множества A к свойству X — это:

$$H(A, X) = -\frac{m}{n}\log_2\frac{m}{n} - \frac{n-m}{n}\log_2\frac{m}{n} \quad (1)$$

Таблица 1. База данных кредитных заявок

Является клиентом > 5 лет?	Отсутствуют ли ДТП?	Стоимость автомобиля >1млн?	Есть ли серьезные нарушения ПДД?	Решение
Нет	Да	Да	Да	Нет
Нет	Да	Да	Нет	Да
Нет	Да	Нет	Нет	Да
Да	Да	Нет	Нет	Да
Да	Нет	Нет	Нет	Нет
Да	Да	Нет	Да	Да
Нет	Нет	Да	Да	Нет
Да	Нет	Да	Нет	-

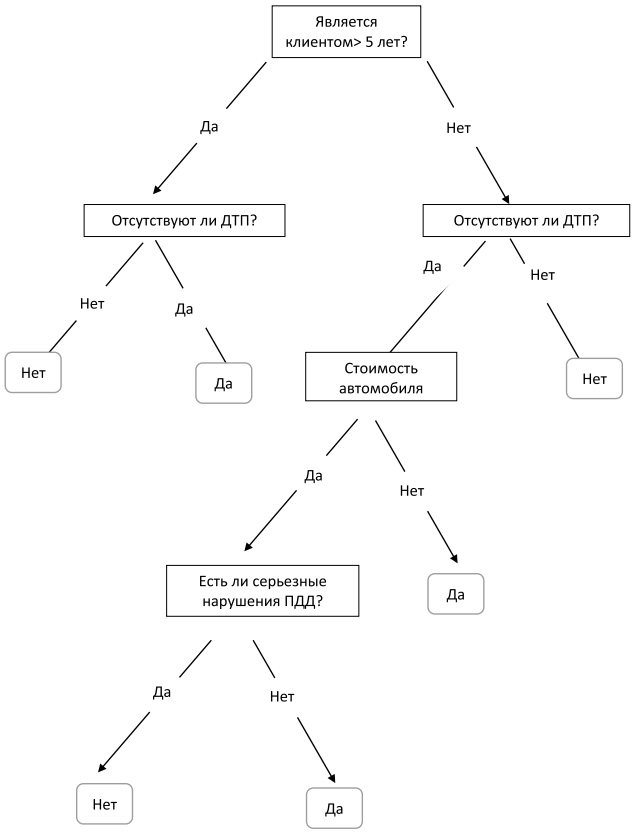

Рис. 1.Фактически составленное дерево

Пусть имеется множество A состоящее из n элементов, у m из которых есть определенное свойство X, принимающее x различных значений, реализуемых в m_i случаях. Тогда энтропия множества A к свойству X — это:

$$H(A,X) = -\sum_{i=1}^{x} \frac{m_i}{n} \log_2 \frac{m_i}{n} \quad (2)$$

Пусть множество A состоящее из элементов, у некоторых из которых имеется определенное свойство X, классифицировано с помощью атрибута Y, у которого есть y значений. Тогда прирост информации или information gain зададим как:

$$Gain(A,Y) = H(A,X) - \sum_{i=1}^{x} \frac{|A_i|}{|A|} H(A_i,X) \quad (3)$$

где A_i — множество элементов A, где X имеет значение i

Сформулируем задачу, которую мы будем рассматривать далее:

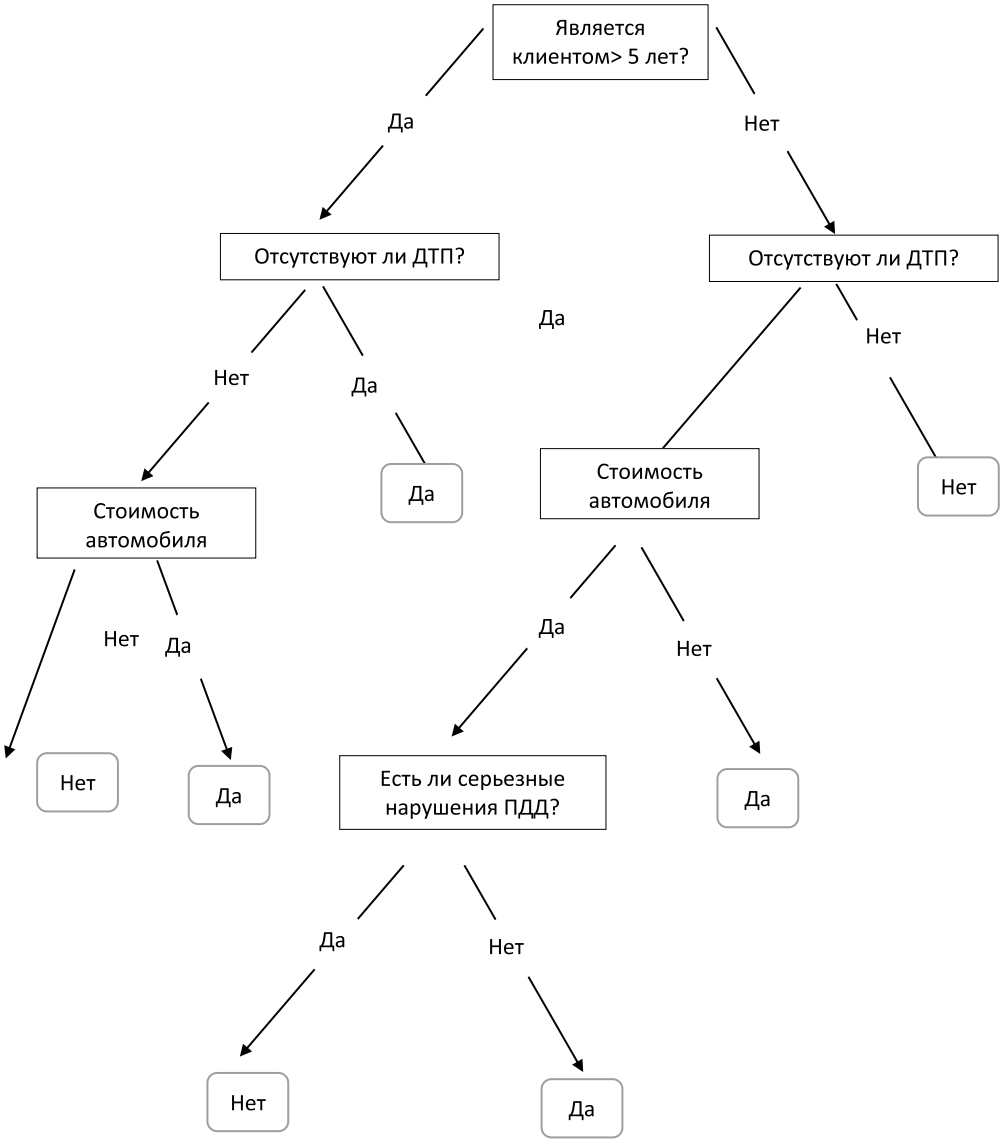

Рис. 2. Фактически составленное дерево, с изменениями

Отделу, страховой компании, по работе с клиентами нужно принять решение о продление страховки на машину. У них имеется некоторая база данных, которая содержит информацию по клиентам и решения по их страховкам.— см. табл. 1.

Нужно понимать, что решение по продлению страховки содержит в себе множество параметров, но мы остановимся на четырех из них.

- Человек является клиентом более 5 лет?
- Являлся ли клиент виновником дорожно-транспортного происшествия?
- Стоимость автомобиля более одного миллиона рублей?
- Имеются ли серьезные нарушения правил дорожного движения?

Новый случай, который ранее еще не встречался в базе данных. То есть человек не являлся клиентом более 5 лет, он являлся виновником ДТП, его автомобиль дороже миллиона рублей и у него нету серьезных нарушений ПДД. Как понять каким же будет решение?

Итак, построим дерево относительно имеющихся данных. Мы пойдем по порядку, то есть поместим в корень информацию о том, как долго клиент является клиентом страховой компании и далее по порядку. В результате получается дерево, изображенное на рис. 1.

Теперь пройдемся по дереву в соответствии с новым случаем. Клиент пользуется услугами компании не более 5 лет, идем налево. Клиент попадал в ДТП, поэтому идем снова налево. По нашему дереву мы получили ре-

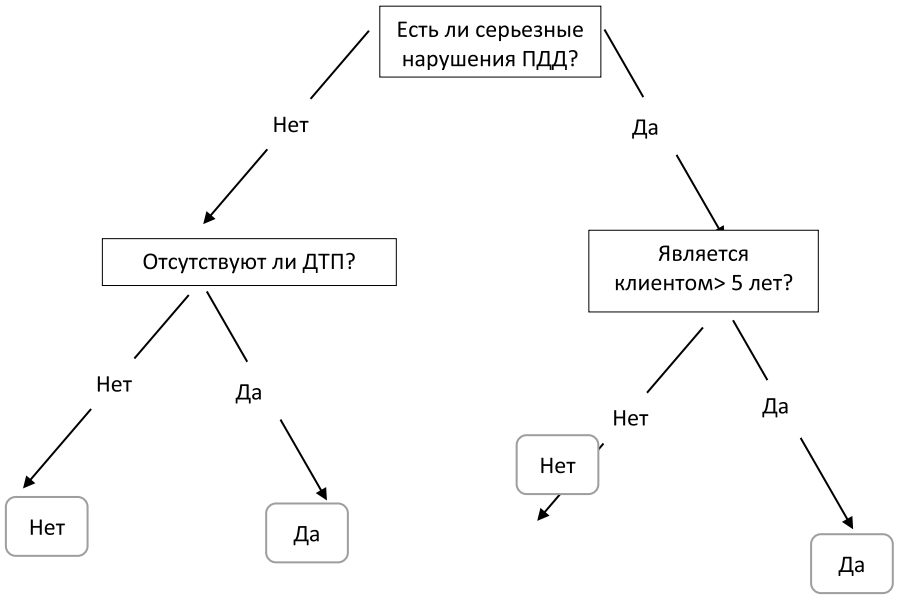

Рис. 3. Оптимальное дерево

1. $Gain(A, \text{явл.клиентом}) = H(A, \text{Одобрено}) - \frac{4}{7}H(A_{\text{нет}}, \text{Одобрено}) - -\frac{3}{7}H(A_{\text{да}}, \text{Одобрено})$

$$Gain(A, \text{явл.клиентом}) = 0{,}98 + \frac{2}{7}(\log_2\frac{1}{2} + \log_2\frac{1}{2}) + \frac{1}{7}(2\log_2\frac{2}{3} + \log_2\frac{1}{3}) = = 0{,}02$$

2. $Gain(A, \text{ДТП}) = H(A, \text{Одобрено}) - \frac{5}{7}H(A_{\text{Да}}, \text{Одобрено}) - -\frac{2}{7}H(A_{\text{Нет}}, \text{Одобрено})$

$$Gain(A, \text{ДТП}) = 0{,}98 - \frac{5}{7}(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}) - \frac{2}{7}(-\frac{2}{2}\log_2\frac{2}{2}) = 0{,}46$$

3. $Gain(A, \text{Цена}) = H(A, \text{Одобрено}) - \frac{3}{7}H(A_{\text{Да}}, \text{Одобрено}) - -\frac{4}{7}H(A_{\text{Нет}}, \text{Одобрено})$

$$Gain(A, \text{Цена}) = 0{,}98 - \frac{3}{7}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{4}{7}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) = 0{,}12$$

4. $Gain(A, \text{ПДД}) = H(A, \text{Одобрено}) - \frac{3}{7}H(A_{\text{Да}}, \text{Одобрено}) - -\frac{4}{7}H(A_{\text{Нет}}, \text{Одобрено})$

$$Gain(A, \text{ПДД}) = 0{,}98 - \frac{3}{7}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{4}{7}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) = 0{,}12$$

Рис. 4

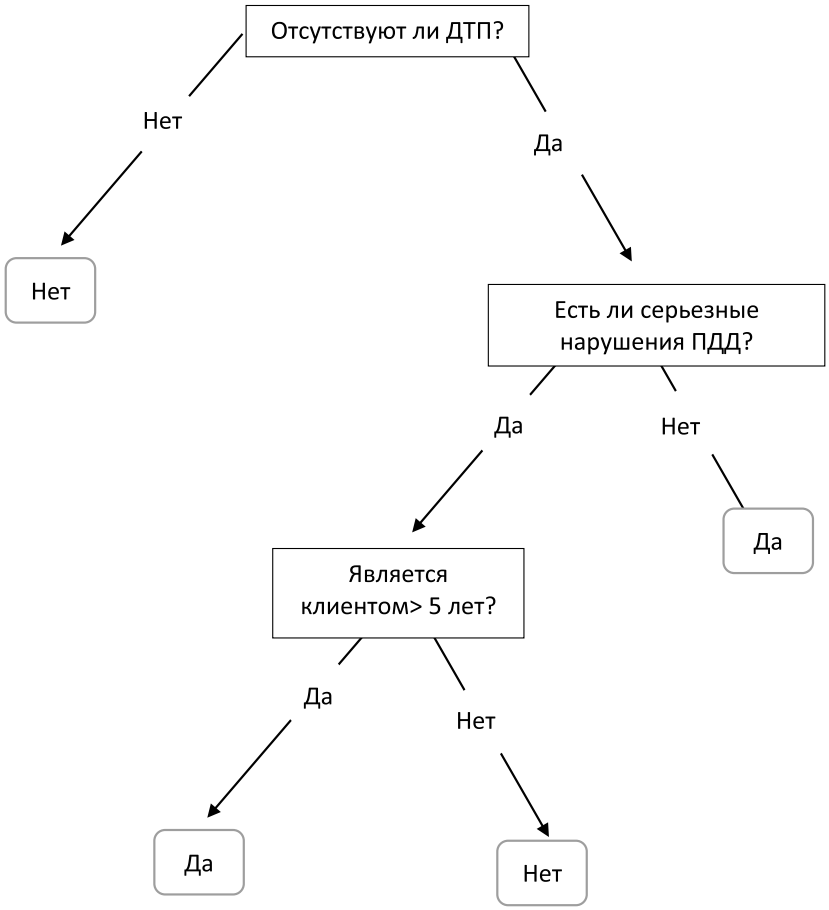

Рис. 5. Оптимизированное дерево с использованием формул

зультат, что в продлении страховки клиенту будет отказано. Отлично мы получили результат и дерево остается прежним, ничего менять не нужно.

Но что, если клиенту продлили страховку, тогда дерево нужно перестроить или вернее дополнить новым узлом и вид его стал таким, как показано на рис. 2.

Является ли это дерево оптимальным? Очевидно, что нет, так как его глубина равна четырем. Для того чтобы сократить глубину можно попробовать взять другой атрибут в корне, например, имеются ли серьезные нарушения ПДД у клиента и посмотрим, что произойдет с нашим деревом рис. 3.

На фоне приведенных данных понятно, чтобы построить оптимальное дерево, нужно выбирать атрибуты, которые лучше всего будут характеризовать нашу функцию. Когда растет от 0 к ½ пропорция, то энтропия растет, а при убывании пропорции от ½, убывает и энтропия. Получается, чтобы выбрать атрибут оптимально, нам требуется чтобы энтропия стала как можно меньше. Энтропия будет разной в разных частях дерева, так что чтобы правильно просуммировать энтропию [2].

Теперь с помощью данных нам формул и определений найдем оптимальный атрибут для корня дерева. Вычислим энтропию для случая, когда нам одобрили продление страховки:

$$H(A, \text{Одобрено}) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \approx 0{,}98$$

Затем вычислим прирост информации для разных случаев (рис. 4).

Получатся, что лучше всего в корень помещать информацию об отсутствие дорожно-транспортных происшествий. Далее будет приведено соответствующее дерево. (рис. 5).

В силу вышеизложенного, становится понятно, чтобы построить оптимальное дерево, нужно выбирать атрибуты, которые лучше всего будут характеризовать нашу функцию. Энтропия зависит от пропорций, как было разделено данное множество. Когда растет от 0 к ½ пропорция, то энтропия растет, а при убывании пропорции от ½, убывает и энтропия. Получается, чтобы выбрать атрибут оптимально, нам требуется чтобы энтропия стала как можно меньше. Энтропия будет разной в разных частях дерева.

1. $Gain(A, \text{явл.клиентом}) = H(A, \text{Одобрено}) - \frac{4}{7}H(A_{\text{нет}}, \text{Одобрено}) - \\ -\frac{3}{7}H(A_{\text{да}}, \text{Одобрено})$

$$Gain(A, \text{явл.клиентом}) = 0{,}98 + \frac{2}{7}(\log_2\frac{1}{2} + \log_2\frac{1}{2}) + \frac{1}{7}(2\log_2\frac{2}{3} + \log_2\frac{1}{3}) = \\ = 0{,}02$$

2. $Gain(A, \text{ДТП}) = H(A, \text{Одобрено}) - \frac{5}{7}H(A_{\text{Да}}, \text{Одобрено}) - \\ -\frac{2}{7}H(A_{\text{Нет}}, \text{Одобрено})$

$$Gain(A, \text{ДТП}) = 0{,}98 - \frac{5}{7}(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}) - \frac{2}{7}(-\frac{2}{2}\log_2\frac{2}{2}) = 0{,}46$$

3. $Gain(A, \text{Цена}) = H(A, \text{Одобрено}) - \frac{3}{7}H(A_{\text{Да}}, \text{Одобрено}) - \\ -\frac{4}{7}H(A_{\text{Нет}}, \text{Одобрено})$

$$Gain(A, \text{Цена}) = 0{,}98 - \frac{3}{7}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{4}{7}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) \\ = 0{,}12$$

4. $Gain(A, \text{ПДД}) = H(A, \text{Одобрено}) - \frac{3}{7}H(A_{\text{Да}}, \text{Одобрено}) - \\ -\frac{4}{7}H(A_{\text{Нет}}, \text{Одобрено})$

$$Gain(A, \text{ПДД}) = 0{,}98 - \frac{3}{7}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{4}{7}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) \\ = 0{,}12$$

Рис. 6

Теперь с помощью данных формул найдем оптимальный атрибут для корня дерева. Вычислим энтропию для случая, когда нам одобрили продление страховки:

$$H(A, \text{Одобрено}) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \approx 0{,}98$$

Как мы получили данное число.

Подставим n=7, m=4 в Опр.2.6.

Затем вычислим прирост информации для разных случаев (рис. 6):

Если же мы решим, после нахождения оптимального корня, найти поэтому же алгоритму оптимальный узел, затем следующий оптимальный узел и так далее, то столкнемся со следующей проблемой. Каждый узел будет оптимальным, но дерево в целом вовсе не обязательно будет оптимальным, а скорее всего его глубина будет велика. В конце концов получится очень подробное дерево, в котором будет возникать огромное число ошибок.

ЛИТЕРАТУРА

1. Т.Коннолли, К. Бегг — Базы данных. Проектирование, реализация и сопровождение. Теория и практика. (3-е издание) — М.: Издательский дом «Вильяме», 2004. — 1440 с.
2. Чубукова И. А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с

Журнал «Современная наука: актуальные проблемы теории и практики»