ПОДХОД К ПОСТРОЕНИЮ КЛАССИФИЦИРУЮЩЕГО ДЕРЕВА НА ВИРТУАЛЬНЫХ ДАННЫХ

Герман Олег Витольдович,
канд. техн. наук, доцент, Белорусский государственный университет информатики и радиоэлектроники, Беларусь, г. Минск

Герман Юлия Олеговна,
канд. техн. наук, доцент, Белорусский государственный университет информатики и радиоэлектроники, Беларусь, г. Минск

Наср Сара Набиб,
аспирантка, Белорусский государственный университет информатики и радиоэлектроники, Беларусь, г. Минск

A NEW APPROACH TO BUILD CLASSIFYING TREE ON VIRTUAL DATA

Oleg German,
candidate of technical sciences, associate Professor, Belarussian state university of informatics and radioelectronics, Republic of Belarus, Minsk

Julia German,
candidate of technical sciences, associate Professor, Belarussian state university of informatics and radioelectronics, Republic of Belarus, Minsk

Sara Nasr,
PhD student, Belarussian state university of informatics and radioelectronics, Republic of Belarus, Minsk

Абстракт

Представлен новый подход к построению иерархического классифицирующего дерева, отличающийся тем, что не требует для создания дерева обучающей таблицы с реальными экспериментальными данными. Вместо этого используется техника полного факторного эксперимента с некоторыми ухищрениями, позволяющими снизить вероятность ложного распознавания практически до нуля (о качестве распознавания). Подробно описаны и теоретически обоснованы все шаги предложенного технического решения, а также доказана теорема об обеспечиваемом качестве распознавания.

Изложение иллюстрируется примером. Результат статьи могут использоваться научными работниками и инженерами при создании систем классификации, кластеризации, прогнозирования и пр.

Ключевые слова: классифицирующее дерево, качество классификации, виртуальные данные.

Abstract

A new approach to construction of a hierarchical classifying tree is presented, which differs in that it does not require a training table with real experimental data for training. Instead, the technique of a full factorial experiment is used with some tweaks to reduce the probability of false recognition to almost zero (the quality of recognition). All steps of the proposed technical solution are described in detail and theoretically substantiated, and a theorem on the quality of recognition provided is proved. The presentation is illustrated by an example. The result of the article can be used by scientists and engineers when creating systems for classification, clustering, forecasting, etc.

Keywords: classifying tree, classification quality, virtual data.

Введение

Применение классифицирующих деревьев сопряжено с двумя интересными задачами: минимизацией числа признаков (характеристик) и минимизацией издержек, связанных со сбором данных для обучения. Что касается первой проблемы, то они так или иначе решаются в существующих подходах к построению иерархических классифицирующих деревьев (ИКД) [1-3], правда вычислительная эффективность различных подходов варьирует в зависимости от требуемой точности. В связи с этим в настоящей статье изложен теоретически оптимальный метод эффективный в среднестатистическом смысле [4]. Весьма интересна вторая проблема, связанная с заменой реальных данных для обучения модельными (как мы говорим – виртуальными). Этим обеспечивается снижение затрат на хранение и

разработку базы данных и сокращение сроков создания системы классификации. Этот пункт составляет основную идею настоящей статьи.

Формализация задачи

В задачах многоальтернативного выбора используют интегральную оценочную функцию вида

$$I = \sum_i \alpha_i f_i(F_i), \quad (1)$$

где α_i определяют веса (приоритеты) критериев F_i (характеристик, features – в англоязычной литературе), представляющих собой неотрицательные вещественные числа в диапазоне [0,1], сумма которых равна 1, и $f_i(F_i)$ обозначают функции полезности критериев. Для отыскания аналитической формы (1) используют, например, метод иерархий Саати [5], метод Relief [6] и его известные модификации и др. Здесь мы полагаем вид (1) известным и используем в качестве примера следующую запись

$$\begin{aligned} I = {} & 0.06 f(F_1) + 0.25 f(F_2) + 0.35 f(F_3) + 0.06 f(F_4) + 0.1 f(F_5) + \\ & 0.1 f(F_6) + 0.1 f(F_7) + 0.04 f(F_8) + 0.04 f(F_9), \end{aligned} \quad (2)$$

Таким образом имеется девять критериев с указанными в (2) весами.

Обычно в задачах классификации (кластеризации, прогнозирования, идентификации и подобных) в качестве исходных данных задается набор векторов (многомерных объектов) *DS* (Data Set) в форме таблицы со строками, представляющими индивидуальные объекты, и столбцами, соответствующими характеристикам (атрибутам) f_i. С помощью формулы (1) *DS* можно разбить на кластеры, например, на два кластера, в первый из которых попадают объекты, для которых значение $I \geq 0.5$, а во второй – те объекты, для которых $I < 0.5$. И вообще, используя (1), проблема отнесения произвольного многомерного объекта в соответствующий кластер становится тривиальной.

Возможна ситуация, когда *DS* не задан изначально, известна лишь интегральная оценка (1) (построение которой, например, по методу Саати не

требует знания *DS*). В этой ситуации все еще актуальна задача построения классифицирующего механизма либо в форме (1), либо в форме классифицирующего дерева [1–3] и др. При этом в записи (1), особенно при большом числе критериев, часть критериев (иногда значительная) может быть избыточной, т.е. не влиять на результат классификации. Таким образом, мы изначально ориентируемся на решение следующих задач в общей связке при исходно заданном функционале *I*:

– найти замену отсутствующему изначально набору данных *DS*;

– минимизировать число критериев за счет отбрасывания максимального числа избыточных критериев F_i.

Решение этих задач дано ниже с использованием конкретного примера (2).

Минимизация числа критериев

Мы будем ориентироваться на получение в конечном итоге иерархического классифицирующего дерева (ИКД), с помощью которого любой входной многомерный объект можно отнести в некоторый результирующий класс.

Введем некоторые базовые идеи [4] и рассмотрим таблицу 1 как пример *DS*, заменяющий реальные данные модельными (объяснение дано далее по тексту).

Таблица 1

Фрагмент модельного набора *DS* к примеру (2)

	$f_1(F_1)$	$f_2(F_2)$	$f_3(F_3)$	$f_4(F_4)$	$f_5(F_5)$	$f_6(F_6)$	$f_7(F_7)$	$f_8(F_8)$	$f_9(F_9)$	I
d_1	0.15	0.85	0.85	0.85	0.15	0.15	0.15	0.15	0.15	0.627
d_2	0.85	0.85	0.15	0.85	0.15	0.15	0.15	0.15	0.15	0.424
d_3	0.85	0.85	0.85	0.85	0.15	0.15	0.15	0.85	0.15	0.697
d_4	0.85	0.15	0.85	0.15	0.85	0.85	0.15	0.85	0.15	0.62

При построении таблицы 1 мы используем формулу (2) для вычисления значений интегральной функции выбора *I*. Для получения данных в столбцах f_1,..., f_9, будем использовать технику полного факторного эксперимента [7]. Согласно этой технике каждая функция полезности для критериев F_i принимает

только два возможных значения: одно – на расстоянии в 15% от минимального значения (т.е. от 0), а второе – на расстоянии в 85% от минимального значения (т.е. 15% от максимального значения, равного 1. Таким образом общее число объектов данных в модельном *DS* составит $2^9 = 512$. В действительности, эту общую идею факторного эксперимента нам нужно скорректировать: вместо значений 0.15 и 0.85 следует сгенерировать два близких случайных числа, скажем 0.149081 и 0.850307. Цель этих действий мы объясним позднее. Пока в рассуждениях будем ориентироваться на 0.15 и 0.85.

Далее нас интересует классификация на основе двух кластеров (классов) *A* и *B*, причем в кластер *A* попадают объекты со значением *I*, большим 0.,5 а в *B* попадают объекты со значением *I*, не превосходящим 0.5.

Определение 1. Характеристика F_t различает два объекта $x \in A$ и $y \in B$, если и только если $f_t(F_{xt}) \neq f_t(F_{yt})$.

Обратим внимание, что данное определение не требует вовсе условие $f_t(F_{xt}) \geq 0$, $f_t(F_{yt}) < 0$, что и является отличительной чертой ИКД.

Определение 2. Множество π характеристик F_i является различающим для данного набора *DS*, если для любых объектов d_i и d_j из *DS*, принадлежащих разным классам, имеется характеристика $F_p \in \pi$, различающая d_i и d_j.

Определение 3. Множество π является минимальным (по числу характеристик) различающим множеством для данного набора *DS* при условии, что не существует различающего множества с меньшим числом характеристик.

Лемма. По отношению к заданной функции выбора *I* из (1) и набору данных *DS* два минимальных различающих множества $\pi(F)$ с элементами $F_1, F_2, ..., F_Z$ и $\pi(f)$ с элементами $f_k(F_k)$, представляющими функции полезности от F_k, $k = 1, z$ имеют одинаковые размеры, то есть $z = Z$, а также находятся во взаимно однозначном соответствии друг с другом.

Доказательство. Пусть для простоты будет только два различных класса *A* и *B*. Пусть F_t различает два объекта d_r и d_s, но f_t их не различает, то есть $F_t(d_r) \neq$

$F_t(d_s)$, но $f_t(F_t(d_r)) = f_t(F_t(d_s))$. Ясно, что должна найтись другая характеристика F_p, различающая d_r и d_s и принадлежащая $\pi(F)$. В самом деле, если нет ни одной такой характеристики, то $I(d_r) = I(d_s)$ и d_r и d_s должны принадлежать одному и тому же классу (кластеру), что невозможно. Поэтому найдется хотя бы одна характеристика $F_q \in \pi(F)$ с $F_q(d_r) \neq F_q(d_s)$ и $f_q(F_q(d_r)) \neq f_q(F_q(d_s))$. Тогда можно включить f_q в $\pi(f)$ и исключить f_t. Эти рассуждения сохраняют силу в отношении любой пары объектов d_r и d_s из различных классов *DS* и показывают как поставить два множества $\pi(F)$ и $\pi(f)$ во взаимно однозначное соответствие, что завершает доказательство.

Отыскание минимального различающего множества

Для отыскания минимального различающего множества характеристик мы используем задачу о минимальном покрытии 0,1-матрицы *M* (называемую далее различающей) множеством строк. Строим различающую матрицу по таблице 1 (исходному набору *DS*) с элементами $m_{kij} = 1$, если и только если f_k различает объекты *i* и *j*; в противном случае $m_{kij} = 0$ (рисунок 1). Строки матрицы соответствуют функциям полезности, столбцы представлены парами (*i*, *j*) где *i* и *j* задают две различные строки в таблице 1. Например, рассмотрим строку f_2 и столбец (1, 2), на пересечении которых стоит «0». Это означает, что f_2 не различает объекты d_1 и d_2. С другой стороны, f_2 различает объекты d_2 и d_4, поскольку $f_2(d_2)=0.85$, но $f_2(d_4)=0.15$. Значения других элементов матрицы *M* получаются по указанному общему правилу.

	(1,2)	(2,3)	(2,4)
f_1	1	0	0
f_2	0	0	1
f_3	1	1	1
f_4	0	0	1
f_5	0	0	1
f_6	0	0	1
f_7	0	0	0
f_8	0	1	1
f_9	0	0	0

Рисунок.1 Различающая 0,1-матрица *M* для табл. 1

Ясно, что среди столбцов различающей матрицы нет тех, которые представляют пары объектов из одного класса. Кроме того, в матрице M не должно быть столбцов, содержащих только нулевые элементы. Кроме того, мы не рассматриваем случай, когда нет характеристик, различающих классы (этот вырожденный случай соответствует недостаточности используемых критериев для построения ИКД). Наша задача, таким образом, свелась к отысканию минимального покрытия матрицы M множеством строк.

Определение 4.Строка k покрывает столбец (i, j) в матрице M, если и только если $m_{kij} = 1$.

Определение 5. Минимальное покрытие $\pi_{min}(f)$ для M содержит минимально возможное число элементов f_i, таких, что каждый столбец матрицы M покрывается хотя бы одним элементом из $\pi_{min}(f)$.

Задача отыскания минимального покрытия $\pi_{min}(f)$ может быть решена, как описано в [4] на основе техники, основанной на применении принципа групповых резолюций (п.г.р.). Этот принцип, дающий эффективные в среднем решения, напоминает логический метод резолюций с той разницей, что в порождении резольвент участвует одновременно более двух родительских дизъюнктов (в общем случае). Детали можно найти в [8,9]. Вернемся к таблице 1. Теоретически она порождает различающую матрицу с $256^2/2 = 32768$ столбцами. Однако, только 512 столбцов остаются уникальными, в то время как остальные 32768 столбцов повторяют какие-то из них. Поэтому размеры M ограничены 9 строками и 512 столбцами. Это количество столбцов теоретически может быть порождено 33 различными объектами данных d_i. Очевидно, матрица M может быть просто сгенерирована программно. Найдя $\pi_{min}(f)$, не представляет труда построить классифицирующее дерево, например с помощью аналитического языка Python.

Экспериментальные и другие теоретически результаты

Эксперименты показали, что для построения дерева классификации необходимо учитывать все 9 признаков. Однако некоторые комбинации

функций могут быть исключены, поскольку некоторые альтернативы получают очень низкие итоговые *I*-оценки (например, 0,3 или ниже). Это позволило сократить набор характеристик до 7 элементов, формирующих минимальное различающее множество.

Основной результат, который мы намереваемся обосновать, связан с правомочностью использования таблицы полного факторного эксперимента вместо таблицы с реальными данными. Мы намереваемся доказать, что минимальное распознающее множество на основе функций полезности из таблицы 1 сохранит способность различать реальные объекты, для которых используется интегральная оценка (2). Допустим противное: найдется реальный объект d_α, принадлежащий классу A, который был отнесен к классу B. Пусть d_α представлен вектором функций полезности: $f_\alpha = \langle f_{\alpha 1}, f_{\alpha 2}, ..., f_{\alpha n} \rangle$. Для этого вектора имеется оценка $I(f_\alpha) = \alpha_1 f_{\alpha 1} + \alpha_2 f_{\alpha 2} + ... + \alpha_n f_{\alpha n}$. Добавим этот объект d_α в таблицу 1 факторного эксперимента в виде отдельной строки (строки d_α). В графе I у него будет стоять величина $I(f_\alpha) \geq 0.5$ (принятая для класса A). Определим по этой расширенной таблице минимальное различающее множество признаков. Можно сообразить, что оно совпадет с предыдущим минимальным различающим множеством признаков для исходной таблицы 1. В самом деле, для всех k $f_{\alpha k} \in \{0.149081, 0.850307\}$ – сгенерированные случайные числа, что позволяет допустить то добавленная строка не совпадет ни в одном столбце ни с одной из имеющихся в таблице 1 строк (вероятность такого совпадения близка к нулю). Поэтому в различающей матрице M в каждой строке в новых столбцах $(1, d_\alpha)$, $(2, d_\alpha)$, ..., (N, d_α), будут стоять одни единицы. Этого условия достаточно, чтобы различающее минимальное покрытие не изменилось. Таким образом, нами доказана следующая

Теорема. Вероятность ложной классификации на основе ИКД, построенного по таблице полного факторного эксперимента с двумя крайними значениями для факторов, представленными случайно сгенерированными числами, близкими к 0.15 и 0.85, стремится к 0.

Доказанная теорема позволяет в практических целях не собирать реальные данные при построении результирующего ИКД, а ограничиться модельными факторными значениями, что экономит время, затрачиваемое на разработку классифицирующей системы с очень низкой вероятностью ложных срабатываний.

Заключение

Основное преимущество описанной техники состоит в экономии затрат памяти, поскольку нет необходимости хранить базу данных со значениями признаков. ИКД обрабатывает вектор нормализованных значений признаков (в диапазоне [0..1]). Для построения ИКД используется факторный эксперимент, в результате которого строится матрица различий, которая используется для нахождения минимального различающего покрытия, содержащего оптимальную коллекцию признаков.

В качестве последнего шага применяется скрипт Python для построения дерева классификации. Варьируя граничный уровень *I* между наборами *A* и *B*, можно представить набор ИКД для максимально возможной фильтрации принятых альтернатив. Эксперты могут протестировать различные модели, представленные уравнением (1), чтобы найти веса характеристик, наиболее соответствующих их предпочтениям.

Список литературы

1. Khalid S., Khalil T., Nasreen S. A survey of feature selection and feature extraction techniques in machine learning // Science and information conference. London UK. 2014. pp. 372–378.
2. Sudrajat R., Irianingsih I., Krisnawan D. Analysis of data mining classification by comparison of C4.5 and ID algorithms. IOP Conference Series: Materials and Engineering. 2017. vol. 166. pp.12–31.
3. Vens C., Stryif J., Shietgat L. et al. Decision trees for hierarchical multi-label classification. Machine Learning. 2008. vol.73., №2. pp. 185–214.

4. German, O.V., Nasr S. New method for optimal feature set reduction. Informatics and automation. SPIRAS Proceedings (St. Petersburg, Russia). 2020. vol.19, №6. pp.1198−1221.
5. Saaty T.L., Vargas L.G. Decision making with the analytic Network Process. Springer. 2013. – 370p.
6. Urbanovicz R.J., Meeker M., Cava V.L. et al. Relief-based feature selection: introduction and review. Journal of biomedical informatics. 2018., vol. 85, pp.189−203.
7. Окунь Я. Факторный анализ. М. Статистика, 1974. – 198с.
8. German J. O. One version of the group resolution principle for discrete optimization. Proc. of Intern. conf. Information Technologies and Systems (ITS) 2020. Minsk, BSUIR, 2020, pp.165−167.
9. Capraro A., Fischetti M. A heuristic method for the set covering problem. Operations Research. 2000. vol.47., №5.