Декабрь, № 8 **Физико-математические науки** **2011**

УДК 519.213

ЮРИЙ ВЛАДИМИРОВИЧ СИДОРОВ
кандидат технических наук, доцент кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет
smalt@drevlanka.ru

ПАВЕЛ ВЛАДИМИРОВИЧ КИРИКОВ
преподаватель кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет
lispad@gmail.com

АЛЕКСАНДР АЛЕКСАНДРОВИЧ РОГОВ
доктор технических наук, профессор, заведующий кафедрой теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет
rogov@psu.karelia.ru

СРАВНЕНИЕ ДЕНДРОГРАММ С РАВНЫМ ЧИСЛОМ ВЕРШИН

В статье рассматривается статистический подход к сравнению расстояний между дендрограммами с равным числом вершин, основанный на ряде классических метрик.

Ключевые слова: расстояния между дендрограммами, вероятностный подход, сравнение расстояний

Один из наиболее популярных алгоритмов кластерного анализа – иерархический алгоритм [1]. Результатом этого алгоритма является дендрограмма. При проведении кластеризации часто приходится рассматривать одни и те же объекты с полным и укрупненным набором признаков [4]. Результаты кластеризации могут в этих случаях отличаться, при этом возникает вопрос о критерии, согласно которому можно решать вопрос о том, значительны или незначительны отличия в результатах. Решение этой задачи и некоторых других требует построения метрики, позволяющей находить расстояния между дендрограммами.

ОПРЕДЕЛЕНИЕ РАССТОЯНИЯ МЕЖДУ ДЕНДРОГРАММАМИ

Пусть A, B – две дендрограммы, содержащие n объектов. Обозначим A^k, B^k – k-е слои данных дендрограмм. Первоначально каждая дендрограмма содержит n объектов, каждый из которых выделен в свой кластер. На каждом слое происходит объединение двух кластеров в один, в результате на k-м слое дендрограмма содержит n – k кластеров. В таком случае функцию схожести $\rho(A, B)$ двух дендрограмм можно задать как функцию от функции схожести $\rho(A^k, B^k)$ одинаковых слоев дендрограмм. В качестве такой функции могут выступать $\rho_m(A,B) = \max_{1<=k<=n-1} \rho(A^k, B^k)$ или $\rho_a(A,B) = \frac{1}{n-1}\sum_{k=1}^{n-1}\rho(A^k, B^k)$. В работе [4] и в данном исследовании используется $\rho_a(A, B)$.

Функцию схожести одинаковых слоев дендрограмм можно задать как функцию от функций схожести $\rho(A_i^k, B_j^k)$ для каждого из n – k кластера этого слоя.

В зависимости от постановки задачи следует применять различные подходы к определению данного расстояния. В случае сравнения деревьев безотносительно к способу нумерации узлов, например в задачах сравнения деревьев семантического описания, для корректности значения данной функции необходимо произвести нормировку значения, учитывать инвариантность к способу нумерации объектов и кластеров. Этим свойствам будет удовлетворять следующая метрика: $\rho(A^k, B^k) = \frac{1}{n-k}\min_{\theta\in\Theta}\sum_{i=1}^{n-k}\rho(A_i^k, B_{\theta_i}^k)$, где $A_i^k, B_{\theta_i}^k$ – соответственно i и θ кластеры k-го слоя дендрограмм A и B, Θ – множество всевозможных n – k элементных размещений номеров групп от 1 до n – k.

При сравнении дендрограмм важно учитывать порядок группировки узлов. В таком случае предлагается применение следующей метрики: $\rho(A^k, B^k) = \frac{1}{n}\sum_{i=1}^{n-k}\mu_i$, $\mu_i = \frac{2 * N_i}{n_{i,1} + n_{i,2}}$ для расстояния Хемминга, $\mu_i = \frac{N_i}{n_{i,1} + n_{i,2} - N_i}$ для расстояния Роджерса – Танимото, где $n_{i,1}$ и $n_{i,2}$ – число объектов в группе, содержащей объект i, соответственно в первом и втором дереве, N_i – число совпадающих элементов в группах, содержащих объект i. В данной статье рассматривается второй подход.

ОПРЕДЕЛЕНИЕ ЗНАЧИМОСТИ РАЗЛИЧИЙ МЕЖДУ ДЕНДРОГРАММАМИ

Независимо от используемых расстояний возникает проблема определения, какие значения расстояний можно считать большими, а какие – малыми для ответа на вопрос о том, является ли отличие между дендрограммами существенным или возникшим случайным образом. В данной статье предлагается подход, основанный на вероятностной модели генерации дендрограмм. Если значение расстояния оказывается

таким, что значения не меньше данного встречаются в рамках модели редко, оно считается «большим» (статистически значимым), а если они встречаются часто – то «малым».

Определим случайный эксперимент, порождающий пару дендрограмм. Расстояние между случайно появившимися дендрограммами будет случайной величиной. На этой основе можно получить вероятностное распределение значений расстояния и провести градуировку интервала возможных значений с помощью квантилей функции распределения расстояния. Пусть λ_a – квантиль уровня a для функции распределения $F_{p(t)}$. Тогда, если расстояние p оказывается не меньше, чем λ_a, можно сделать вывод, что не менее чем a * 100 % случайно выбранных пар разбиений имеют между собой расстояние меньше, чем p. Аналогичный подход рассмотрен авторами в работе [1] при сравнении расстояний между подмножествами.

ВЕРОЯТНОСТНАЯ МОДЕЛЬ ПОЯВЛЕНИЯ ЗНАЧЕНИЯ РАССТОЯНИЯ

Представим кластера $A^k{}_i$ и $B^k{}_j$ в виде бинарных векторов a и b размерности n, построенных по принципу: $a_i = 1$ тогда и только тогда, когда объект u_i входит в кластер A: $u_i \in A$, в противном случае $a_i = 0$ (аналогично для b и B). Обозначим через p_i, $i = 1, \ldots, n$ вероятность появления элемента u_i в кластере. Рассмотрим случайный эксперимент, состоящий из n независимых испытаний, в каждом из которых элемент может как появиться, так и не появиться в кластерах. Тогда в каждом испытании возможны исходы четырех видов $A^i{}_{uv} = \{x_i = u, y_i = v\}$, где $u, v \in \{0,1\}$, i – номер испытания. Пусть $I(A)$ – индикатор события A, $I(A^i_{11}) + I(A^i_{10}) + I(A^i_{01}) + I(A^i_{00}) = 1$, $a = \sum_{i=1}^{n} I(A^i_{11})$, $b = \sum_{i=1}^{n} I(A^i_{10})$, $c = \sum_{i=1}^{n} I(A^i_{01})$, $d = \sum_{i=1}^{n} I(A^i_{00})$. Тогда $a + b + c + d = n$.

Согласно мультиномиальному (полиномиальному) распределению [5], функция распределения расстояния примет вид:

$$F_\rho(t) = P(\rho\ (X,Y) < t) = \sum_{(a,b,c,d)\in C(A^1_{11},\ldots,A^n_{00})} \sum_{B} \prod_{i=1}^{n} p_i^{2I(A^i_{11})} (1-p_i)^{2I(A^i_{00})} (p_i(1-p_i))^{I(A^i_{01})+I(A^i_{10})}, \quad (1)$$

где $C = \{(a,b,c,d) \in Z^4 : a,b,c,d \geq 0,\ a + b + c + d = n, h(a,b,c,d) < t\}$.

Объединение любых двух кластеров на каждом шаге является равновозможным. Будем считать равновозможной нумерацию кластеров на каждом слое. Вычислим вероятность появления произвольного объекта в любом кластере. Вероятность $P^k_{i,j}$ попадания i-го объекта в j-й кластер на k-м слое не зависит от i и j и равняется $\frac{1}{n-k}$.

Учитывая, что $p_i = \frac{1}{n-k}$, формула (1) примет вид:

$$F_\rho(t) = \sum_{(a,b,c,d)\in C(A^1_{11},\ldots,A^n_{00})} \sum_{B} \prod_{i=1}^{n} \left(\frac{1}{n-k}\right)^{2I(A^i_{11})} \left(\frac{n-k-1}{n-k}\right)^{2I(A^i_{00})} \left(\frac{1}{n-k}\left(\frac{n-k-1}{n-k}\right)\right)^{I(A^i_{01})+I(A^i_{10})} = \quad (2)$$

$$= \sum_{(a,b,c,d)\in C(A^1_{11},\ldots,A^n_{00})} \sum_{B} \prod_{i=1}^{n} \frac{1}{(n-k)^{2a}} \frac{(n-k-1)^{2d}}{(n-k)^{2d}} \frac{(n-k-1)^{b+c}}{(n-k)^{2b+2c}} = \sum_{(a,b,c,d)\in C(A^1_{11},\ldots,A^n_{00})} \sum_{B} \prod_{i=1}^{n} \frac{(n-k-1)^{n+d-a}}{(n-k)^{2n}}.$$

РАССТОЯНИЕ ХЕММИНГА

Для конкретных расстояний формула (2) может упрощаться. В [6] в качестве простейшего коэффициента различия между множествами, обладающего свойствами метрики, предлагается мощность симметрической разности: $|X \Delta Y| = |(X \setminus Y) \cup (Y \setminus X)|$. Если разделить это число на n, то получим расстояние Хэмминга [3] между бинарными векторами, принимающее значения от 0 до 1:

$\rho^H(X,Y) = \frac{|X \Delta Y|}{n} = \frac{m}{n}$, где $m = b + c$.

Вероятность исходов A^i_{10} и A^i_{01} равняется $\frac{1}{n-k} * (1 - \frac{1}{n-k}) = \frac{n-k-1}{(n-k)^2}$. Тогда

$$F^H_\rho(t) = P(\rho^H(X,Y) < t) = \sum_{m:\frac{m}{n}<t} C^m_n \frac{2^{b+c}(n-k-1)^{b+c}}{(n-k)^{2b+2c}} \times \times \frac{((n-k)^2 - 2(n-k-1))^{a+d}}{(n-k)^{2a+2d}} =$$

$$= \sum_{m:\frac{m}{n}<t} C^m_n \frac{2^{b+c}(n-k-1)^{b+c}}{(n-k)^{2b+2c}} * \frac{2^{b+c}(n-k-1)^{b+c}}{(n-k)^{2b+2c}}, \quad (3)$$

$$\sum_{m:\frac{m}{n}<t} C^m_n \frac{2^m(n-k-1)^m(n-k)^{2a+2d} - 2^n(n-k-1)^n}{(n-k)^{2n}} =$$

$$= \sum_{m:\frac{m}{n}<t} C^m_n \frac{2^m(n-k-1)^m}{(n-k)^{2m}} - \sum_{m:\frac{m}{n}<t} C^m_n \frac{2^n(n-k-1)^n}{(n-k)^{2n}}.$$

РАССТОЯНИЕ РОДЖЕРСА – ТАНИМОТО

Одним из часто используемых коэффициентов различия между бинарными векторами является расстояние Роджерса – Танимото, получаемое из одноименной меры близости [5] и равное

$$\rho^{RT}(X,Y) = 1 - \frac{a+d}{a+d+2(b+c)} = \frac{2(b+c)}{n+b+c} = \frac{2m}{n+m},$$

где m – количество исходов $A_{01} \cup A_{10}$ в n испытаниях, то есть $m = b + c$.

Число исходов m рассчитывается аналогично предыдущему пункту, следовательно,

$$F_{\rho}^{RT}(t) = P\left(\rho^{RT}(X,Y) < t\right) = \sum_{m:\frac{2m}{n+m}<t} C_n^m \frac{2^m (n-k-1)^m}{(n-k)^{2m}} - \sum_{m:\frac{2m}{n+m}<t} C_n^m \frac{2^n (n-k-1)^n}{(n-k)^{2n}} \sum_{m:\frac{2m}{n+m}<t} C_n^m . \quad (4)$$

ЧИСЛЕННОЕ ВЫЧИСЛЕНИЕ КВАНТИЛЕЙ РАСПРЕДЕЛЕНИЯ

Для подсчета квантилей функций распределения расстояний была разработана специальная компьютерная программа. Результаты работы этой программы представлены в табл. 1, содержащей в себе квантили уровня a для двух рассмотренных в статье расстояний, рассчитанные для различных n. Для каждого a в табл. 1 представлены 2 квантиля, расположенные следующим образом: вверху – квантиль для расстояния Хэмминга, внизу – для расстояния Роджерса – Танимото.

Таблица 1

Квантили функции распределения для различных n (с точностью 0,001)

k\p	0,6	0,7	0,8	0,9	0,95	0,99
3	0,334 0,501	0,334 0,501	0,334 0,501	0,667 0,801	0,667 0,801	0,667 0,801
4	0,334 0,501	0,334 0,501	0,334 0,501	0,667 0,801	0,667 0,801	0,667 0,801
5	0,401 0,572	0,401 0,572	0,401 0,572	0,601 0,751	0,601 0,751	0,801 0,889
15	0,334 0,501	0,334 0,501	0,401 0,572	0,467 0,637	0,534 0,696	0,601 0,751
30	0,334 0,501	0,367 0,537	0,401 0,572	0,434 0,605	0,467 0,637	0,534 0,696
45	0,356 0,525	0,356 0,525	0,378 0,549	0,423 0,594	0,445 0,616	0,489 0,657
60	0,351 0,519	0,367 0,537	0,384 0,555	0,417 0,589	0,434 0,605	0,484 0,652

АНАЛИЗ ДЕНДРОГРАММ ПРИ КЛАССИФИКАЦИИ ТЕКСТОВ С РАЗНЫМ НАБОРОМ ПРИЗНАКОВ

Задача сравнения дендрограмм была поставлена в [4], где анализировались результаты кластеризации объектов с разным количеством признаков. При проведении исследований по атрибуции текстов использовалось: распределение 16 частей речи на первых трех и последних трех позициях каждого предложения. Таким образом, каждому тексту соответствовал набор 96 (16 x 6) признаков; расширенный набор признаков за счет учета дополнительных морфологических характеристик, свойственных каждой части речи (например, падеж для существительных, форму и степень сравнения для прилагательных, вид, залог и лицо для глаголов и т. д.). В итоге стали использоваться 156 признаков, соответственно, текст стал характеризоваться набором в 936 (156 x 6) признаков.

Была обнаружена визуальная схожесть дендрограмм, полученных при иерархической классификации 60 текстов. Проверим эту схожесть с помощью предложенного метода. Коэффициенты близости между иерархическими деревьями приведены в табл. 2.

Таблица 2

Коэффициенты близости между иерархическими деревьями, построенными на основе различного числа признаков: 16 (1) и 156 (2)

По всему тексту	(1)	0,877995
	(2)	0,839487
По первому и последнему предложению абзаца	(1)	0,936056
	(2)	0,914381
По первому предложению абзаца	(1)	0,910971
	(2)	0,880166
По последнему предложению абзаца	(1)	0,941005
	(2)	0,921237

Из табл. 2 видно, что коэффициенты близости текстов принимают значения немногим меньше 1. Сравнивая их с экспериментальными результатами (табл. 1), можно сделать вывод, что значения мер близости встречаются менее чем в 1 % случаев (уровень надежности – 0,99). Это означает, что увеличение числа грамматических признаков не позволяет эффективно решать задачу атрибуции анонимных и псевдонимных литературных произведений. Проведенное исследование показало статистически незначимое отличие классификации текстов по двум наборам признаков.

СПИСОК ЛИТЕРАТУРЫ

1. Варфоломеев А. А., Кириков П. В., Рогов А. А. Вероятностный подход к сравнению расстояний между подмножествами конечного множества // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2010. № 8 (113). С. 83–88.
2. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977. 128 с.
3. Орлов А. И. Нечисловая статистика. М.: МЗ-Пресс, 2004. 513 с.
4. Сидоров Ю. В. Математическая и информационная поддержка методов обработки литературных текстов на основе формально-грамматических параметров: Автореф. дис. … канд. техн. наук. Петрозаводск, 2002. 19 с.
5. Ширяев А. Н. Вероятность. М.: Наука, 1980. 576 с.
6. Rogers D., Tanimoto T. A computer program for classifying plants // Science. 1960. Vol. 132. № 3434. P. 1115–1118.