УЧЕНЫЕ ЗАПИСКИ ПЕТРОЗАВОДСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Декабрь, № 8 **Физико-математические науки** **2010**

УДК 519.213

АЛЕКСЕЙ ГЕННАДЬЕВИЧ ВАРФОЛОМЕЕВ
кандидат физико-математических наук, доцент кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет
avarf@psu.karelia.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

ВЕРОЯТНОСТНЫЙ ПОДХОД К СРАВНЕНИЮ РАССТОЯНИЙ МЕЖДУ ПОДМНОЖЕСТВАМИ КОНЕЧНОГО МНОЖЕСТВА

В статье рассматривается вероятностный подход к сравнению расстояний между подмножествами одного множества, основанный на ряде классических метрик. Получены функции распределения для каждого из приведенных расстояний, приведена таблица квантилей распределений.

Ключевые слова: расстояния между подмножествами, вероятностный подход, сравнение расстояний, бинарные вектора

1. ВВЕДЕНИЕ

В приложениях часто возникают задачи, связанные с вычислением степени различия двух подмножеств одного конечного множества. Типичная ситуация – попарные сравнения кластеров, полученных в результате двух процедур автоматической классификации некоторого множества объектов [12], [16]. Сравнение строк или столбцов матрицы значений дихотомических признаков для объектов, то есть бинарных векторов одинаковой размерности, также можно рассматривать как сравнение двух подмножеств одного множества. Кроме того, на оценке степени различия множеств основаны многие алгоритмы сравнения более сложных объектов, таких как деревья и графы [10], [15].

В качестве степени различия двух подмножеств часто используют расстояние, то есть числовую функцию $\rho = \rho(X,Y)$, заданную на множестве пар подмножеств и обладающую свойствами метрики. Значения расстояния могут расти неограниченно, но часто они ограничены сверху единицей. Это характерно, например, для расстояний, которые получаются из некоторых мер близости [4], [6], [8] *S(X,Y)* с помощью операции дополнения $\rho(X,Y) = 1 - S(X,Y)$.

Независимо от того, каким интервалом ограничены возможные значения расстояния, возникает проблема определения, какие значения можно считать большими, а какие – малыми. Решение данной проблемы могло бы помочь ответить на вопрос о том, является ли отличие между множествами существенным или нет. В данной статье предлагается подход, основанный на вероятностной модели генерации подмножеств. При этом если значение расстояния оказывается таким, что значения не меньше данного встречаются в рамках модели редко, оно считается «большим», а если такие значения встречаются часто, то «малым». В работе предлагается способ построения количественных оценок для понятий «редко» и «часто» на основе вероятностного распределения значений расстояния.

Определим случайный эксперимент, порождающий пару подмножеств, и введем вероятностную меру на множестве исходов эксперимента. На этой основе можно получить вероятностное распределение значений расстояния и провести градуировку интервала возможных значений с помощью квантилей функции распределения расстояния. Пусть λ_α – квантиль уровня α для функции распределения $F_\rho(t) = P(\rho < t)$. Тогда, если расстояние ρ оказывается не меньше, чем λ_α, можно сделать вывод, что не менее чем $\alpha \cdot 100\ \%$ случайно выбранных пар подмножеств имеют между собой расстояние меньше, чем ρ.

В статье рассмотрена одна из возможных моделей генерации подмножеств. Основное внимание уделяется случаю равновероятных подмножеств, который рассмотрен для четырех расстояний – Хэмминга, Роджерса – Танимото, Жаккара и Сокала – Снита. В каждом из этих случаев получена функция распределения расстояния. В заключительном разделе приведена таблица, позволяющая сравнивать значения

рассмотренных расстояний для различных множеств.

2. ВЕРОЯТНОСТНОЕ РАСПРЕДЕЛЕНИЕ РАССТОЯНИЯ – ОБЩИЙ СЛУЧАЙ

Пусть $U=\{u_1,u_2,\ldots,u_n\}$ – конечное множество из n элементов, X и Y – два его подмножества. Представим подмножества X и Y в виде бинарных векторов x и y размерности n, построенных по принципу: $x_i=1$ тогда и только тогда, когда $u_i\in X$, в противном случае $x_i=0$ (аналогично для y и Y). Обозначим через $p_i, i=1,\ldots,n$ вероятность появления элемента u_i в подмножестве. Тогда можно рассмотреть случайный эксперимент, состоящий из n независимых испытаний, в каждом из которых элемент u_i может как появиться, так и не появиться в подмножествах. 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,

$$I\left(A^i_{11}\right)+I\left(A^i_{10}\right)+I\left(A^i_{01}\right)+I\left(A^i_{00}\right)=1,$$

$$a=\sum_{i=1}^{n}I\left(A^i_{11}\right),\ b=\sum_{i=1}^{n}I\left(A^i_{10}\right),\ c=\sum_{i=1}^{n}I\left(A^i_{01}\right),$$

$$d=\sum_{i=1}^{n}I\left(A^i_{00}\right).$$

Тогда $a+b+c+d=n$.

Введем множество событий

$$B=\left\{\left(A^1_{11},A^1_{10},\ldots,A^n_{00}\right):\sum_{i=1}^{n}I\left(A^i_{11}\right)=a,\sum_{i=1}^{n}I\left(A^i_{10}\right)=\right.$$

$$\left.=b,\sum_{i=1}^{n}I\left(A^i_{01}\right)=c,\sum_{i=1}^{n}I\left(A^i_{00}\right)=d\right\}.$$

Во многих работах (см., например, [8], [16]) разные коэффициенты различия между множествами описываются как функции от чисел a, b, c, d, то есть $\rho(X,Y)=h(a,b,c,d)$. В данной работе мы ограничиваемся этим же случаем. Тогда функцию распределения случайной величины $\rho(X,Y)$ можно записать в следующем виде:

$$F_\rho(t)=P\left(\rho\ (X,Y)<t\right)=$$

$$=\sum_{(a,b,c,d)\in C}\ \sum_{\left(A^1_{11},\ldots,A^n_{00}\right)\in B}\ \prod_{i=1}^{n}p_i^{2I\left(A^i_{11}\right)}\times$$

$$\times\left(1-p_i\right)^{2I\left(A^i_{00}\right)}\left(p_i\left(1-p_i\right)\right)^{I\left(A^i_{01}\right)+I\left(A^i_{10}\right)},$$

где

$$C=\left\{(a,b,c,d)\in Z^4:a,b,c,d\ge 0,a+b+c+d=\right.$$

$$\left.=n,h(a,b,c,d)<t\right\}.$$

Рассмотрим важный частный случай. Предположим, что $p_i=0.5, i=1,\ldots,n$, тогда каждое подмножество множества U может появиться в ролях X и Y с одинаковой вероятностью. Согласно мультиномиальному (полиномиальному) распределению [9],

$$F_\rho(t)=P\left(\rho\ (X,Y)<t\right)=$$

$$=\sum_{(a,b,c,d)\in C}\frac{n!}{a!b!c!d!}\cdot\frac{1}{4^a}\cdot\frac{1}{4^b}\cdot\frac{1}{4^c}\cdot\frac{1}{4^d}. \qquad (1)$$

Для конкретных расстояний формула (1) может упрощаться. Рассмотрим несколько расстояний между подмножествами, принимающих значения от 0 до 1. В дальнейшем изложении символ $|X|$ означает мощность множества X. Нам также потребуются две функции вещественного аргумента: $\lfloor t\rfloor=\max\{k\in Z\mid k\le t\}$, $\lceil t\rceil=\min\{k\in Z\mid k\ge t\}$.

3. РАССТОЯНИЕ ХЭММИНГА

В книге [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 [4] между бинарными векторами, принимающее значения от 0 до 1:

$$\rho^H(X,Y)=\frac{|X\Delta Y|}{n}=\frac{m}{n}, \qquad (2)$$

где $m=b+c$.

Легко понять, что случайная величина m имеет биномиальное распределение, так как мощность симметрической разности равна числу исходов $A_{01}\cup A_{10}$ в n испытаниях (вероятность такого исхода в одном испытании равна 1/2). Тогда

$$F^H_\rho(t)=P\left(\rho^H(X,Y)<t\right)=$$

$$=\frac{1}{2^n}\sum_{m:\frac{m}{n}<t}C_n^m=\frac{1}{2^n}\sum_{m=0}^{\lceil tn\rceil-1}C_n^m. \qquad (3)$$

Для нахождения квантилей функции распределения случайной величины $\rho^H(X,Y)$ можно воспользоваться таблицами квантилей биномиального распределения, находящимися, например, в [5]. При $n>36$ в [1] рекомендуют использовать нормальную аппроксимацию, которая в нашем случае будет иметь вид

$$F^H_\rho(t)=P\left(\rho^H(X,Y)<t\right)=P\left(n\rho^H(X,Y)<nt\right)\approx$$

$$\approx\frac{1}{2}+\Phi\left(\frac{2nt+1-n}{\sqrt{n}}\right),$$

где $\Phi(t)=\frac{1}{\sqrt{2\pi}}\int_0^t e^{-(\tau^2/2)}d\tau$ – функция Лапласа.

Максимальная абсолютная ошибка приближенного равенства меньше $0,28/\sqrt{n}$ [1].

4. РАССТОЯНИЕ РОДЖЕРСА – ТАНИМОТО

Одним из часто используемых коэффициентов различия между бинарными векторами является расстояние Роджерса – Танимото, получаемое из одноименной меры близости [13] 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 $m=b+c$.

Очевидно, что

$$\rho^{RT}=\frac{2\rho^H}{\rho^H+1}.$$

Поскольку функция $g(t)=\frac{2t}{t+1}:[0,1]\to[0,1]$ монотонно возрастающая и вогнутая, $\rho^{RT}(X,Y)$ обладает свойствами метрики [2], поэтому имеет право называться расстоянием.

Случайная величина m имеет такое же распределение, что и в п. 1. Следовательно,

$$F_\rho^{RT}(t)=P\left(\rho^{RT}(X,Y)<t\right)=$$
$$=\frac{1}{2^n}\sum_{m:\frac{2m}{n+m}<t}C_n^m=\frac{1}{2^n}\sum_{m=0}^{\left\lceil\frac{tn}{2-t}\right\rceil-1}C_n^m=P\left(m<\frac{tn}{2-t}\right).$$

Для нахождения квантиля функции распределения можно воспользоваться рекомендациями, приведенными в п. 1. При больших n можно использовать нормальную аппроксимацию

$$F_\rho^{RT}(t)\approx\frac{1}{2}+\Phi\left(\frac{3tn+2-t-2n}{(2-t)\sqrt{n}}\right).$$

В этом случае максимальная абсолютная ошибка приближенного равенства меньше $0,28/\sqrt{n}$ [1].

5. РАССТОЯНИЕ ЖАККАРА

Во многих публикациях также рассматривается расстояние между множествами в виде отношения мощности симметрической разности к мощности их объединения. Свойства метрики для него доказаны, например, в [7]. Расстояние получается из коэффициента близости П. Жаккара [11], предложенного в 1901 году для сравнения двух географических областей по населяющим их биологическим видам. Аналогичная мера предложена в [15] для измерения расстояния между графами. Итак,

$$\rho^J(X,Y)=\frac{|X\Delta Y|}{|X\cup Y|}=1-\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|}=\frac{b+c}{b+c+a},\quad(4)$$

если $X\cap Y\neq\varnothing$; $\rho^J(\varnothing,\varnothing)=1$.

Пусть по-прежнему $m=b+c$. Тогда

$$\rho^j(X,Y)=m/(m+a).$$

Согласно полиномиальному распределению:

$$F_\rho^J(t)=P\left(\rho^J(X,Y)<t\right)=$$
$$=\sum_{(a,m)\in B}C_n(a,m)\cdot\frac{1}{4^a}\cdot\frac{1}{2^m}\cdot\frac{1}{4^{n-a-m}}=$$
$$=\sum_{(a,m)\in B}\frac{n!\cdot 2^m}{a!m!(n-a-m)!\cdot 4^n},\quad(5)$$

где

$$B=\left\{(a,m)\in Z^2:a,m\geq 0,\frac{m}{a+m}<t,0<a+m\leq n\right\}.$$

Заметим, что двойная сумма (5) может быть записана как повторная:

$$F_\rho^J(t)=\sum_{m=0}^{n}\sum_{a=\left\lfloor\frac{m(1-t)}{t}\right\rfloor+1}^{n-m}C_n(a,m)\cdot\frac{1}{4^a}\cdot\frac{1}{2^m}\cdot\frac{1}{4^{n-a-m}}=$$
$$=\sum_{m=0}^{n}C_n^m\frac{2^m}{4^n}\sum_{a=\left\lfloor\frac{m(1-t)}{t}\right\rfloor+1}^{n-m}C_{n-m}^a.$$

Функция распределения $F_\rho^J(t)=P\left(\rho^J(X,Y)<t\right)$ случайной величины $\rho^j(X,Y)$ не протабулирована в справочниках по математической статистике. Поэтому в статье приводится краткая таблица некоторых значений квантилей этой функции (табл. 1).

В случае большого n для вычисления слагаемых в формуле (5) можно воспользоваться локальной теоремой Муавра – Лапласа для полиномиального распределения [3]. В нашем случае равенство примет вид:

$$\frac{n!}{m!a!(n-m-a)}\frac{2^m}{4^n}\approx\frac{2\sqrt{2}}{\pi n}e^{-\frac{3x^2+4xy+4y^2}{n}},\quad(6)$$

где $x=m-\frac{n}{2}$, $y=a-\frac{n}{4}$.

Таблица 1

Таблица квантилей функции распределения расстояния Жаккара

n\α	0,01	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,99
1	0,001	0,001	0,001	0,001	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
2	0,001	0,001	0,001	0,501	0,501	0,501	1,000	1,000	1,000	1,000	1,000	1,000	1,000
3	0,001	0,001	0,001	0,334	0,501	0,667	0,667	1,000	1,000	1,000	1,000	1,000	1,000
4	0,001	0,001	0,334	0,501	0,501	0,667	0,667	0,751	1,000	1,000	1,000	1,000	1,000
5	0,001	0,251	0,334	0,501	0,501	0,601	0,667	0,751	0,801	1,000	1,000	1,000	1,000
6	0,001	0,251	0,334	0,501	0,501	0,601	0,667	0,751	0,801	0,834	1,000	1,000	1,000
7	0,167	0,334	0,401	0,501	0,572	0,601	0,667	0,751	0,801	0,834	1,000	1,000	1,000
8	0,167	0,334	0,401	0,501	0,572	0,626	0,667	0,715	0,801	0,834	1,000	1,000	1,000
9	0,201	0,334	0,429	0,501	0,572	0,626	0,667	0,715	0,751	0,834	0,876	1,000	1,000
10	0,251	0,376	0,429	0,501	0,572	0,626	0,667	0,715	0,751	0,834	0,876	1,000	1,000
11	0,251	0,376	0,445	0,501	0,572	0,626	0,667	0,715	0,751	0,801	0,876	0,901	1,000
12	0,273	0,401	0,445	0,546	0,601	0,626	0,667	0,715	0,751	0,801	0,876	0,901	1,000
13	0,286	0,401	0,455	0,546	0,601	0,637	0,667	0,701	0,751	0,801	0,876	0,901	1,000
14	0,301	0,417	0,462	0,546	0,601	0,637	0,667	0,701	0,751	0,801	0,847	0,901	1,000
15	0,308	0,417	0,501	0,546	0,601	0,637	0,667	0,701	0,751	0,801	0,847	0,901	1,000
16	0,334	0,429	0,501	0,546	0,601	0,637	0,667	0,701	0,751	0,786	0,834	0,901	1,000
17	0,334	0,445	0,501	0,546	0,601	0,637	0,667	0,701	0,751	0,786	0,834	0,876	0,934
18	0,358	0,455	0,501	0,563	0,601	0,643	0,667	0,701	0,734	0,778	0,834	0,867	0,934
19	0,358	0,462	0,501	0,563	0,601	0,643	0,667	0,706	0,734	0,770	0,824	0,867	0,934
20	0,364	0,462	0,501	0,563	0,601	0,643	0,667	0,706	0,734	0,770	0,824	0,867	0,934
21	0,376	0,467	0,501	0,563	0,601	0,643	0,667	0,706	0,734	0,770	0,819	0,858	0,934
22	0,385	0,467	0,501	0,563	0,612	0,643	0,667	0,706	0,734	0,765	0,813	0,858	0,929
23	0,389	0,471	0,527	0,572	0,612	0,643	0,667	0,701	0,734	0,765	0,813	0,847	0,924
24	0,392	0,474	0,527	0,572	0,612	0,648	0,667	0,701	0,728	0,765	0,810	0,843	0,905
25	0,401	0,477	0,527	0,579	0,612	0,648	0,667	0,701	0,728	0,762	0,801	0,843	0,901
26	0,410	0,479	0,527	0,579	0,612	0,648	0,667	0,701	0,723	0,762	0,801	0,843	0,901
27	0,412	0,501	0,527	0,579	0,612	0,648	0,667	0,701	0,723	0,761	0,801	0,834	0,901
28	0,417	0,501	0,527	0,579	0,612	0,643	0,667	0,696	0,723	0,751	0,801	0,834	0,895
29	0,422	0,501	0,530	0,584	0,616	0,641	0,667	0,696	0,723	0,751	0,801	0,827	0,895
30	0,424	0,501	0,542	0,584	0,616	0,641	0,667	0,696	0,721	0,751	0,792	0,827	0,885
31	0,429	0,501	0,542	0,584	0,616	0,641	0,667	0,696	0,721	0,751	0,792	0,827	0,881
32	0,435	0,501	0,542	0,584	0,616	0,643	0,667	0,696	0,721	0,751	0,792	0,822	0,881
33	0,435	0,501	0,542	0,587	0,620	0,643	0,667	0,693	0,721	0,751	0,786	0,819	0,876
34	0,441	0,501	0,542	0,587	0,620	0,643	0,667	0,693	0,721	0,751	0,786	0,819	0,876
35	0,445	0,518	0,546	0,591	0,621	0,643	0,667	0,693	0,715	0,751	0,786	0,815	0,871
36	0,445	0,518	0,549	0,593	0,621	0,643	0,667	0,693	0,715	0,742	0,783	0,815	0,870
37	0,449	0,518	0,552	0,593	0,621	0,643	0,667	0,693	0,715	0,742	0,782	0,813	0,864
38	0,452	0,518	0,552	0,593	0,621	0,646	0,667	0,690	0,715	0,742	0,778	0,808	0,863
39	0,455	0,518	0,552	0,593	0,621	0,646	0,667	0,690	0,715	0,742	0,778	0,808	0,863
40	0,460	0,519	0,552	0,594	0,621	0,646	0,667	0,690	0,715	0,741	0,778	0,807	0,858

6. РАССТОЯНИЕ СОКАЛА – СНИТА

В монографии [14] были введены несколько коэффициентов близости бинарных векторов, один из которых имеет вид

$$\frac{a}{a+2(b+c)}.$$

Если применить к этому коэффициенту операцию дополнения, описанную в п. 1, то получится коэффициент различия

$$\rho^{SS}(X,Y)=\frac{2(b+c)}{a+2(b+c)}.$$

Нетрудно показать, что $\rho^{SS}=g(\rho^{J})$, где $g(t)$ – функция, определенная в п. 4. Следовательно, $\rho^{SS}(X,Y)$ – расстояние на множестве подмножеств фиксированного множества. Функция распределения $F_{\rho}^{SS}(t)=P(\rho^{SS}(X,Y)<t)$ определяется по формуле (5) с той только разницей, что множество B имеет несколько иной вид:

$$B=\left\{(a,m)\in Z^2 : a,m\geq 0, \frac{2m}{a+2m}<t, 0<a+m\leq n\right\}.$$

Аналогично предыдущему пункту для облегчения вычислений можно использовать ап-

проксимацию (6), а также запись функции распределения через повторную сумму:

$$F_{\rho}^{SS}(t)=\sum_{m=0}^{n}\sum_{a=\left\lfloor\frac{m(2-t)}{t}\right\rfloor+1}^{n-m}C_n(a,m)\cdot\frac{1}{4^a}\cdot\frac{1}{2^m}\cdot\frac{1}{4^{n-a-m}}=$$

$$=\sum_{m=0}^{n}C_n^m\frac{2^m}{4^n}\sum_{a=\left\lfloor\frac{m(2-t)}{t}\right\rfloor+1}^{n-m}C_{n-m}^a.$$

Для получения квантилей для функции распределения расстояния Сокала – Снита можно также воспользоваться табл. 1. Действительно, если λ_α – квантиль уровня α для случайной величины ξ, а $\tilde{\lambda}_\alpha$ – квантиль того же уровня для случайной величины $\zeta=\varphi(\xi)$, где $\varphi(t)$ – монотонная функция вещественного аргумента, то указанные квантили связаны соотношением $\tilde{\lambda}_\alpha=\varphi(\lambda_\alpha)$. Поэтому квантиль λ_α^{SS} может быть найден как $g\left(\lambda_\alpha^J\right)$, где $g(t)$ – функция, определенная в п. 4.

7. СРАВНЕНИЕ РАССТОЯНИЙ

Предложенная во введении идея сравнения расстояний позволяет сравнивать значения расстояния, полученные на разных множествах и с помощью разных метрик. Такая задача возникает, например, при кластеризации одного набора объектов с разным набором признаков.

Для двух разных расстояний ρ^1 и ρ^2, двух подмножеств X_1, Y_1 одного множества и двух подмножеств X_2, Y_2 другого множества будем считать значения расстояний $\rho^1(X_1,Y_1)$ и $\rho^2(X_2,Y_2)$ одинаковыми, если они равны квантилям функций распределения расстояний одного и того же уровня α. Если же уровни α_1 и α_2 различны, но отличаются друг от друга на величину, не превосходящую ε, то будем называть такие значения расстояний близкими с точностью ε.

Найдем вероятность того, что значения функций распределений расстояний $\rho^1(X_1,Y_1)$ и $\rho^2(X_2,Y_2)$ друг от друга отличаются на ε. Она будет равна

$$P\left(\left|F_{\rho^1}(\rho^1(X_1,Y_1))-F_{\rho^2}(\rho^2(X_2,Y_2))\right|<\varepsilon\right)=$$

$$=P\left(F_{\rho^1}(\rho^1(X_1,Y_1))-\varepsilon<F_{\rho^2}(\rho^2(X_2,Y_2))<\right.$$

$$\left.<F_{\rho^1}(\rho^1(X_1,Y_1))+\varepsilon\right)=2\varepsilon$$

при условии, что $F_{\rho^1}(\rho^1(X_1,Y_1))-\varepsilon>0$ и $F_{\rho^1}(\rho^1(X_1,Y_1))+\varepsilon<1$.

Таблица 2

Сравнительная таблица квантилей функций распределения четырех расстояний

α\n	5		7		10		15		20		30		40	
0.01	0.001	0.001	0.143	0.251	0.101	0.182	0.201	0.334	0.251	0.401	0.301	0.462	0.326	0.491
	0.001	0.001	0.167	0.286	0.251	0.401	0.308	0.471	0.364	0.534	0.424	0.595	0.46	0.63
0,05	0.201	0.334	0.143	0.251	0.201	0.334	0.267	0.422	0.301	0.462	0.367	0.537	0.376	0.546
	0.251	0.401	0.334	0.501	0.376	0.546	0.417	0.589	0.462	0.632	0.501	0.667	0.519	0.683
0,1	0.201	0.334	0.286	0.445	0.301	0.462	0.334	0.501	0.351	0.519	0.367	0.537	0.401	0.572
	0.334	0.501	0.401	0.572	0.429	0.601	0.501	0.667	0.501	0.667	0.542	0.703	0.552	0.712
0,2	0.401	0.572	0.286	0.445	0.401	0.572	0.401	0.572	0.401	0.572	0.434	0.605	0.426	0.597
	0.501	0.667	0.501	0.667	0.501	0.667	0.546	0.706	0.563	0.721	0.584	0.737	0.594	0.746
0,3	0.401	0.572	0.429	0.6	0.401	0.572	0.401	0.572	0.451	0.621	0.467	0.637	0.451	0.621
	0.501	0.667	0.572	0.728	0.572	0.728	0.601	0.751	0.601	0.751	0.616	0.762	0.621	0.766
0,4	0.401	0.572	0.429	0.6	0.501	0.667	0.467	0.637	0.451	0.621	0.467	0.637	0.476	0.645
	0.601	0.751	0.601	0.751	0.626	0.77	0.637	0.778	0.643	0.783	0.641	0.781	0.646	0.785
0,5	0.401	0.572	0.429	0.6	0.501	0.667	0.467	0.637	0.501	0.667	0.501	0.667	0.501	0.667
	0.667	0.801	0.667	0.801	0.667	0.801	0.667	0.801	0.667	0.801	0.667	0.801	0.667	0.801
0,6	0.601	0.751	0.572	0.728	0.501	0.667	0.534	0.696	0.551	0.71	0.534	0.696	0.526	0.689
	0.751	0.858	0.751	0.858	0.715	0.834	0.701	0.824	0.706	0.828	0.696	0.821	0.69	0.817
0,7	0.601	0.751	0.572	0.728	0.601	0.751	0.601	0.751	0.551	0.71	0.534	0.696	0.551	0.71
	0.801	0.889	0.801	0.889	0.751	0.858	0.751	0.858	0.734	0.847	0.721	0.838	0.715	0.834
0,8	0.601	0.751	0.715	0.834	0.601	0.751	0.601	0.751	0.601	0.751	0.567	0.724	0.576	0.731
	1	1	0.834	0.91	0.834	0.91	0.801	0.889	0.77	0.87	0.751	0.858	0.741	0.852
0.9	0.801	0.889	0.715	0.834	0.701	0.824	0.667	0.801	0.651	0.788	0.634	0.776	0.601	0.751
	1	1	1	1	0.876	0.934	0.847	0.917	0.824	0.904	0.792	0.884	0.778	0.876
0.95	0.801	0.889	0.858	0.924	0.801	0.889	0.734	0.847	0.701	0.824	0.634	0.776	0.626	0.77
	1	1	1	1	1	1	0.901	0.948	0.867	0.929	0.827	0.905	0.807	0.893
0.99	1	1	0.858	0.924	0.901	0.948	0.801	0.889	0.751	0.858	0.701	0.824	0.676	0.806
	1	1	1	1	1	1	1	1	0.934	0.966	0.885	0.939	0.858	0.924

Полученное утверждение можно использовать для построения статистического критерия близости значений расстояний.

Продемонстрируем предложенный нами метод сравнения расстояний с помощью табл. 2, содержащей в себе квантили уровня α для четырех рассмотренных в статье расстояний, рассчитанные при различных n. Для каждой пары α и n в табл. 2 представлены 4 квантиля, расположенные в виде мини-таблицы 2 x 2 следующим образом: в первой строке слева – квантиль для расстояния Хэмминга, справа – для расстояния Роджерса – Танимото, во второй строке слева – квантиль для расстояния Жаккара, справа – для расстояния Сокала – Снита. Интерпретация табл. 2 такова: значения соответствующих расстояний для одного и того же α мы предлагаем считать одинаковыми (например, значение 0,429 для расстояния Жаккара при n = 10 и значение 0,537 для расстояния Роджерса – Танимото при n = 30 одинаковы), значения же расстояний при разных α – близкими с точностью, равной модулю разности значений α.

СПИСОК ЛИТЕРАТУРЫ

1. Вадзинский Р. Н. Справочник по вероятностным распределениям. СПб.: Наука, 2001. 295 с.
2. Деза М., Лоран М. Геометрия разрезов и метрик. М.: МЦНМО, 2001. 736 с.
3. Калинин В. М. Специальные функции и предельные свойства вероятностных распределений. II // Записки научных семинаров ЛОМИ. 1972. Т. 26. С. 5–87.
4. Мандель И. Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.
5. Мюллер П., Нойман Н., Шторм Р. Таблицы по математической статистике. М.: Финансы и статистика, 1988. 272 с.
6. Орлов А. И. Нечисловая статистика. М.: МЗ-Пресс, 2004. 513 с.
7. Петровский А. Б. Пространства множеств и мультимножеств. М., 2003. 248 с.
8. Раушенбах Г. В. Меры близости и сходства // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 169–203.
9. Ширяев А. Н. Вероятность. М.: Наука, 1980. 576 с.
10. Bunke H., Shearer K. A graph distance metric based on the maximal common subgraph // Pattern Recognition Letters. 1998. Vol. 19. P. 255–259.
11. Jaccard P. The distribution of the flora in the Alpine zone // New Phytologist. 1912. Vol. 11. Issue 2. P. 37–50.
12. Meila M. Comparing Clusterings by the Variation of Information // Learning Theory and Kernel Machines. Lecture Notes in Computer Science. Vol. 2777. Springer. 2003. P. 173–187.
13. Rogers D., Tanimoto T. A computer program for classifying plants // Science. 1960. Vol. 132. № 3434. P. 1115–1118.
14. Sokal R. R., Sneath P. H. Principles of numerical taxonomy. San Francisco: W. H. Freeman and Company, 1963.
15. Wallis W. D., Shoubridge P., Kraetzl M., Ray D. Graph distances using graph union // Pattern Recognition Letters. 2001. Vol. 22. P. 701–704.
16. Warrens M. J. Similarity Coefficients for Binary Data: Dissertation. Leiden, 2008.