ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 20 МАТЕМАТИКИ Выпуск 4
2013

М. П. Кривенко (Москва, ИПИ РАН)**. Стратификация данных о химическом составе камней при уролитиазе.**

Клинические и метаболические характеристики мочекаменной болезни могут существенно изменяться с течением времени [2]. Исследование соответствующих реальных клинических данных позволяет формально обосновать данное предположение, а также понять истинную значимость наработок в области прикладной математики и информатики.

Материалом для исследования служили результаты анализа химического состава более 4 000 удаленных оперативно мочевых конкрементов, при этом рассматривались задачи анализа зависимости данных о составе камней от пола пациента и времени. Существенная повторяемость значений признаков и невозможность использования модели нормального распределения вынуждают обратиться к непараметрическим методам. Пусть имеются выборки объема m и n, представляющие собой два множества точек в евклидовом пространстве, тогда в качестве статистики для проверки гипотезы об однородности берется расстояние между этими множествами. Справедливо утверждение [1]: для независимых случайных векторов X_1, X_2, Y_1, Y_2, где X_1, X_2 имеют одно то же распределение F с конечным значением $\mathbf{E}\{\|X_1\|\}$ и Y_1, Y_2 — G с конечным значением $\mathbf{E}\{\|Y_1\|\}$, действует неравенство

$$\mathbf{E}\{\|X_1 - Y_1\|\} - \frac{1}{2}\mathbf{E}\{\|X_1 - X_2\|\} - \frac{1}{2}\mathbf{E}\{\|Y_1 - Y_2\|\} \geqslant 0,$$

которое становится равенством тогда и только тогда, когда $F = G$. Замена левой части неравенства ее выборочным аналогом приводит к статистике $T_{m,n}$ для проверки нулевой гипотезы $H : F = G$, против конкурирующей $K : F \neq G$. Для получения критических уровней значимости приходится прибегать к непараметрическому бутстреп-методу.

Для обнаружения зависимости состава камней от пола пациента критерий $T_{m,n}$ используется напрямую, в случае же зависимости от времени – как основа для парных сравнений фрагментов данных, относящихся к определенному промежутку времени (было выделено 5 фрагментов, соответствующих одному году обследования). Анализ показал, что 5-й фрагмент значимо отличается от остальных, а первые 4 фрагмента дают не совсем ясную картину.

Установим для пары фрагментов отношение совпадения распределений данных, оно будет выполняться, если при проверке гипотезы H критический уровень значимости будет превосходить некоторый порог. Введенное отношение не обязательно является транзитивным. Но транзитивность вместе с имеющимися рефлексивностью и симметрией приводят к отношению эквивалентности и, следовательно, к возможности получить разбиение для исходного множества фрагментов — стратифицировать совокупность данных.

Поставим задачу нахождения транзитивного приближения, наиболее близкого к заданному отношению. Если R_1 и R_2 — матрицы рефлексивных, симметричных отношений, то для определения их меры близости используем расстояние Хемминга, а именно: $\rho(R_1, R_2) = |\{i, j : 1 < i < j < k, R_1(i, j) \neq R_2(i, j)\}|$. Тогда, если R_0 — матрица заданного отношения, то искомой матрицей $R_T^*(R_0)$ из множества $\mathfrak{R}_T$ матриц всех рефлексивных, симметричных, транзитивных отношений будет $R_T^*(R_0) = \arg(\min_{R \in \mathfrak{R}_T} \rho(R_0, R))$. Для исследуемых данных разбиение, соответствующее $R_T^*(R_0)$, есть $\{\{1, 2, 3, 4\}, \{5\}\}$. С его помощью можно не только обоснованно провести разбиение исходных данных, но и объяснять содержание отличий фрагментов.

СПИСОК ЛИТЕРАТУРЫ

1. *Baringhaus L., Franz C.* On a new multivariate two-sample test. — J. of Multivariate Analysis, 2004, v. 88, p. 190-206.
2. *Trinchieri A., Coppi F., Montanari E., Del Nero A., Zanetti G., Pisani E.* Increase in the Prevalence of Symptomatic Upper Urinary tract Stones during the Last Ten Years. — Eur. Urol., 2000, v. 37, p. 23–25.