$$Q(k)=\{(x_p,y_p)\,|\,(x_p-x_i^{расч})^2+(y_p-y_i^{расч})^2=d_i^2;$$
$$x_i-k\overline{\varepsilon_{x_i}}\le x_i^{расч}\le x_i+k\overline{\varepsilon_{x_i}};y_i-k\overline{\varepsilon_{y_i}}\le y_i^{расч}\le y_i+k\overline{\varepsilon_{y_i}};i\in\overline{1,n}\},$$

где $\overline{\varepsilon_{x_i}}=\overline{\varepsilon_{y_i}}$ – оценки модулей ошибок в расчетных значениях $x_i^{расч}$, $y_i^{расч}$. которые принадлежат прямоугольнику со сторонами $2\overline{\varepsilon_{x_i}}=2\overline{\varepsilon_{y_i}}$, $i\in\overline{1,n}$.

Применение данной методики позволяет нивелировать ошибки изменений исходных пунктов на начальной стадии.

Библиографический список

1. Инженерная геодезия: учебник для вузов / Е.Б.Клюшкин, М.И. Киселев, Д.Ш. Михеев [и др.]; под ред. Д.Ш. Михеева. – М. : Высш. шк., 2000. – 464 с.

2. Суханов В.А. Исследование эмпирических зависимостей: нестатистический подход: сборник научных статей / под. ред. Н.М. Оскорбина. П.И. Кузьмина. – Барнаул : Изд-во АлтГУ. 2007. – 290 с.

УДК 519.24

О максимально различных кластерных разбиениях конечного множества

А.П. Фоменко

АлтГУ, г. Барнаул

При первичной обработке больших массивов данных специалист в любой отрасли науки чаще всего вынужден прибегать к разбиению их на примерно однородные группы, что можно назвать кластеризацией данных. В силу этого кластерные алгоритмы и методы сегодня актуальны и активно разрабатываются, см. [1]. Тем не менее, практически всеми практиками признается, что довольно часто объективная кластеризация невозможна, т.е. при применении разных алгоритмов к одному и тому же множеству объектов в итоге могут получиться существенно разные его разбиения. При этом, изучая степень максимально возможного различия двух кластерных разбиений заданного множества объектов, можно, например, делать заключения о возможности его «объективно правильной» кластеризации.

Целью работы является получение неулучшаемой оценки наибольшего возможного различия кластерных разбиений некоторого множества U из n объектов, если известны количества кластеров в каждом из этих разбиений и использование этой оценки для определения нового вида связи между нечисловыми категорированными признаками.

Под кластерным разбиением условимся понимать систему непустых дизъюнктных подмножеств множества U, объединение которых совпадает с U. Обычно объекты из одного кластера считаются близкими в каком-то смысле. Ниже мы будем строить кластеры, как наборы объектов, которые обладают одной и той же категорией некоторого нечислового признака, и в этом смысле являются близкими. Именно такое понимание и оправдывает то, что элементы нашей системы подмножеств мы далее будем называть кластерами, а само в достаточной степени произвольное разбиение кластерным.

На семействе кластерных разбиений множества U, следуя [2], введем расстояние (метрику). Пусть изучаемые разбиения A и B состоят из кластеров $A_1,...,A_k$ и $B_1,...,B_m$ соответственно. Обозначим количества элементов в их попарных пересечениях $a_{i.j}=|A_i\cap B_j|,\quad i=1,...,k;\, j=1,...,m$. Поместим все числа $a_{i.j}$ в матрицу размерности $k\times m$, которую назовем матрицей пересечений. Нам потребуется также число

$$T_{i,j}=\sum_{t\neq j}a_{i,t}+\sum_{s\neq i}a_{s,j},\quad i=1,...,k;\, j=1,...,m,$$

которое равно числу элементов соответственной симметрической разности кластеров. Тогда вводимое в [2] расстояние вычисляется по формуле

$$d(A,B)=\sum_{i,j}a_{i,j}T_{i,j}\,. \tag{1}$$

Пусть зафиксировано некоторое кластерное разбиение основного множества U, которое далее будем называть базовым. Допустим, что в нем m кластеров, а количество элементов i-кластера обозначим через x_i , i= 1,…, m. Ясно, что все x_j натуральные числа, а их сумма равна общему числу объектов n.

Условимся разбиения, состоящие из заданного количества r кластеров, называть r-разбиениями. Пусть задано натуральное число k. Поставим задачу среди всех k-разбиений найти то, для которого значение расстояния d от базового m-разбиением является наибольшим. Для этого будем строить матрицу пересечений двух разбиений.

Путем исследования величины (1) на экстремальные значения при наборе условий

$$\sum_{j=1}^{k}a_{i,j}=x_i,\quad i=1,...m, \tag{2}$$

удалось доказать, что справедливо следующее утверждение.

Теорема 1. *Если число элементов x_i в каждом кластере базового разбиения кратно k, то максимальное расстояние d от базового m-разбиения до некоторого k-разбиения достигается, когда каждая строка матрицы пересечений содержит лишь равные между собой числа. Значение этого максимума задается формулой*

$$d_{\max} = \frac{n^2}{k} + \frac{k-2}{k}\sum_{i=1}^{m} x_i^2 .$$

Видим, что d_{max} зависит от чисел x_j и, следовательно, изменяя количества элементов в кластерах базового разбиения, даже не меняя количества этих кластеров, мы будем получать разные значения соответствующего максимума. При этом

$$\sum_{i=1}^{m} x_i = n, \quad (\forall i)(\exists z_i \in N)\ x_i = kz_i .$$

Экстремальные значения суммы квадратов таких чисел и условия для их достижения приведены в [3]. Отсюда вытекает

Теорема 2. *В условиях теоремы 1 наибольшее и наименьшее из значений d_{max} достигаются, когда все строки матрицы пересечений содержат лишь равные между собой элементы. При этом наибольшее из них равно*

$$\overline{d} = \left((n-(m-1)k)^2 + (m-1)k^2\right)\cdot\frac{k-2}{k} + \frac{n^2}{k}.$$

и получается, когда все кластеры m-разбиения, кроме одного, содержат по k элементов. Наименьшее d_{max} достигается, если каждый из кластеров m-разбиения относится к одной из двух групп, причем в каждой из групп кластеры одинаковы по числу элементов, но каждый кластер одной из групп содержит ровно на k объектов больше, каждый кластер другой. Это наименьшее значение d_{max} равно

$$\underline{d} = \frac{n^2}{k} + n(k-2)\cdot\left(2\left[\frac{n}{km}\right]+1\right) - mk(k-2)\cdot\left[\frac{n}{km}\right]\cdot\left(\left[\frac{n}{km}\right]+1\right).$$

Чтобы наглядно представить себе утверждение теоремы 1, рассмотрим семейства всех k-разбиений $\boldsymbol{U_k}$ и m-разбиений $\boldsymbol{U_m}$ множества U как два непересекающихся многообразия в пространстве всех возможных его кластерных разбиений. Тогда те из базовых разбиений, которые обладают наибольшим d_{max} можно считать образующими внешнюю границу $\boldsymbol{U_m}$, а те из k-разбиений, которые максимально от них удалены, образующими противоположный участок внешней границы $\boldsymbol{U_k}$. Те же из m-разбиений, которые соответствуют минимальному d_{max}, лежат как бы в центре семейства $\boldsymbol{U_m}$. При этом максимально удаленные от них k-разбиения расположены на границах многообразия $\boldsymbol{U_k}$.

Если отбросить требование делимости каждого из x_j на k, и, тем самым, разрешить m-разбиению быть в большей степени произвольным, то имеет место следующее утверждение.

Теорема 3. *Наибольшее расстояние d между k- и m-разбиениями основного множества из n элементов удовлетворяет неравенству*

$$d_{\max} \le \frac{(k+m-2)n^2}{km}.$$

Верхняя граница в этом неравенстве достигается, если n делится нацело на произведение km.

Пусть показатель X имеет m категорий, а показатель Y – k категорий. Возьмем обучающую выборку U из n объектов, у каждого из которых известна категория как X, так и Y. Тогда можно рассмотреть два разбиения A_X, A_Y множества U: в каждом из кластеров A_X содержатся элементы, попадающие в одну категорию по X, а в A_Y кластеры составлены из элементов с одинаковой категорией по Y. Будем говорить, что показатели кластерно d-связаны, если расстояние d между этими кластерными разбиениями имеет достаточно малую величину (разбиения похожи).

Тогда можно ввести принимающее в [0,1] число

$$J(X,Y) = \frac{d(A_X, A_Y)\cdot km}{(k+m-2)\cdot n^2},$$

и назвать его коэффициентом кластерной d-связи. Чем величина коэффициента $J(X,Y)$ меньше, тем сильнее эта степень.

Результаты работы в частном случае k=m=2 были успешно применены автором к задаче выявления значимых признаков ошибочного соединения в сессиях отправки данных с привлечением в качестве обучающей выборки набора данных KDD Cup 1999.

Библиографический список

1. Chance B.L., Rossman A.J. Investigating Statistical Concepts // Applications, and Methods. – Duxbury Press, 2013.

2. Дронов С.В. Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. – 2011. – Вып. 1 / 2 (69).

3. Dronov S.V., Evdokimov E.A. Post-hoc cluster analysis of connection between the forming characteristics // Model Assisted Statistics and Applications. – 2018. – № 2.