М.В. Хачумов

Расстояния, метрики и кластерный анализ[1]

Аннотация. Дается обзор некоторых метрик для решения задач классификации и кластеризации. Приведены определения и теоремы. Рассмотрены вопросы оптимизации первоначального размещения кластеров. Проведены эксперименты по равномерному размещению кластеров в задаче коммивояжера.

Ключевые слова: метрики, кластеризация, классификация, начальное размещение кластера, эксперименты

Введение

В задачах кластерного анализа и классификации часто используются расстояния, основанные на функции Махаланобиса [1,2], с помощью которых можно определять сходство образов и классов. Они отличается от расстояния Евклида тем, что учитывают дисперсии переменных, называемых также признаками. Функцию, предложенную Махаланобисом, традиционно называют метрикой, что не достаточно очевидно и требует дополнительного исследования. Кластеризация предназначена для объединения объектов в непересекающиеся группы (кластеры, классы, множества) «близких» объектов, каждый из которых характеризуется набором признаков. Типовым является подход, связанный с выбором в качестве первоначальных кластеров заданного числа случайных представителей из множества имеющихся экземпляров, подлежащих кластеризации. Однако, как показывают эксперименты, это часто приводит к увеличению числа итераций и не дает хороших результатов. Имеются предложения по равномерному распределению объектов-кластеров по m-сфере [3,4]. Целью работы является, выбор адекватной меры, необходимой для решения задач кластеризации и классификации, и обоснование одной из задач первоначального размещения заданного числа кластеров.

1. Расстояния и метрики. Основные определения и понятия

Под метрикой, как правило, понимают функцию или формулу, определяющую расстояние между любыми точками и классами в метрическом пространстве R^p [5]. Метрическое пространство есть множество точек с функцией расстояния d. В литературе, таким образом, не делается разницы между расстоянием и метрикой, что вносит некоторую нечеткость в понятиях. Далее рассмотрим определения расстояния и метрики, предложенные в работе [6]. Пусть задано произвольное подмножество $X \subseteq R^p$.

Определение 1. Функция $d : X \times X \to R$ называется ***расстоянием***, если она удовлетворяет следующим условиям:

[1] Работа выполнена при поддержке РФФИ (проект №09-07-00006-а), программы фундаментальных исследований ОНИТ РАН «Информационные технологии и методы анализа сложных систем» (проект 2.2); ФЦП «Научные и научно-педагогические кадры инновационной России" (лот шифр: 2010-1.1-411-009), тема: «Разработка технологии интеллектуальной обработки информации в командно-измерительных системах космического назначения» (шифр: 2010-1.1-411-009-033); Программы № 17 фундаментальных исследований Президиума РАН (проект «Разработка инструментальных программных средств обработки потоков изображений высокого разрешения и данных широкого назначения с применением суперкомпьютерных систем»).

а) неотрицательность: $d(x,y) \geq 0$ для всех $x, y \in X$;

б) идентичность (совместимость): $d(x,y) = 0$, если и только если $x = y$;

в) симметричность: $d(x,y) = d(y,x)$ для всех $x, y \in X$.

Рассмотрим любые три p- мерные точки $x, y, z \in X$. Расстояние $d(x,y)$ называется ***метрикой,*** если выполняется следующее дополнительное условие:

г) неравенство треугольника: $d(x,y) \leq d(x,z) + d(z,y)$.

Определение 2. Евклидовым расстоянием (Euclidian Distance) между двумя точками $x = (x_1,...,x_p)^T$ и $y = (y_1,...,y_p)^T$ в пространстве R^p называется функция вида [1]:

$$d_E(x,y) = \sqrt{(x_1 - y_1)^2 + ... + (x_p - y_p)^2} = \\ = \sqrt{(x-y)^T (x-y)}$$

При этом $d_E(x,0) = \|x\|_2 = \sqrt{x_1^2 + ... + x_p^2} = \sqrt{x^T x}$ является евклидовой нормой x. Из этого необходимо следует, что все точки с одним и тем же расстоянием, имеющим норму $d_E(x,0) = \|x\|_2 = c$, удовлетворяют равенству $x_1^2 + ... + x_p^2 = c^2$, которое определяет уравнение сфероида.

Утверждение 1. Расстояние Евклида является **метрикой** в пространстве R^p.

Доказательство этого утверждения, выполненное на основе неравенства Коши–Буняковского, приведено, например, в работе [7].

В задачах кластеризации предпочитают изменение каждого из признаков определять отклонением от центра, причем признаки с высокой изменчивостью должны получить меньший вес, чем признаки с низкой изменчивостью. Это может быть достигнуто путем масштабирования. Нормализованное расстояние применяют также для признаков, измеренных в различных единицах или существенно различающихся по величине. Пусть

$$u = \left(\frac{x_1}{s_1},...,\frac{x_p}{s_p}\right),\ v = \left(\frac{y_1}{s_1},...,\frac{y_p}{s_p}\right),$$

тогда нормализованное расстояние Евклида между точками x и y вычисляется следующим образом:

$$d(x,y) = d_E(u,v) = \sqrt{\left(\frac{x_1 - y_1}{s_1}\right)^2 + ... + \left(\frac{x_p - y_p}{s_p}\right)^2} = \\ = \sqrt{(x-y)^T D^{-1} (x-y)},$$

где: $D = diag(s_1^2,...,s_p^2)$, s_i –масштабирующие коэффициенты. Теперь $\|x\| = d(x,0) = d_E(u,0) =$

$$= \|u\|_2 = \sqrt{\left(\frac{x_1}{s_1}\right)^2 + ... + \left(\frac{x_p}{s_p}\right)^2} = \sqrt{x^T D^{-1} x}$$

и все точки с одинаковым расстоянием с нормой $\|x\| = c$ удовлетворяют уравнению $\left(\frac{x_1}{s_1}\right)^2 + ... + \left(\frac{x_p}{s_p}\right)^2 = c^2$, которое является уравнением эллипсоида.

Определение 3. Статистическим расстоянием или расстоянием Махаланобиса (Mahalanobis Distance) между двумя точками $x = (x_1,...,x_p)^T$ и $y = (y_1,...,y_p)^T$ в пространстве R^p называют функцию вида [1]:

$$d_M(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)} \qquad (1)$$

и $d_M(x,0) = \|x\|_S = \sqrt{x^T S^{-1} x}$ является нормой x. Здесь S - матрица ковариации. Расстояние Махаланобиса можно определить как меру несходства между двумя случайными векторами $x = (x_1,...,x_p)^T$ и $y = (y_1,...,y_p)^T$ из одного распределения вероятностей с матрицей ковариации S. Пусть $z_i = (z_{i1},...,z_{ip})^T$ и $z_j = (z_{j1},...,z_{jp})^T$ – два вектора-строки размерности p (т.е. $z_i \in R^p$ и $z_j \in R^p$ для всех i, j). Матрица ковариации размерности $p \times p$ для N точек определяется как:

$$S = \frac{1}{N-1} A^T A,$$

где: $A = \begin{pmatrix} (z_{11},...,z_{1p})^T - (z_{01},...,z_{0p})^T \\ (z_{21},...,z_{2p})^T - (z_{01},...,z_{0p})^T \\ ... \\ (z_{N1},...,z_{Np})^T - (z_{01},...,z_{0p})^T \end{pmatrix}$.

Здесь $(z_{01},...,z_{0p})^T$ – точка, относительно которой измеряется расстояние. В дальнейшем под центром класса будем понимать точку, определяемую средними значениями параметров, т.е. $(\bar{z}_1,...,\bar{z}_p)^T$. Элемент матрицы ковариации S вычисляют следующим образом: $\sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(z_{ki} - \bar{z}_i)(z_{kj} - \bar{z}_j);\ i,j = 1,...,p$.

Все точки с одним и тем же расстоянием, имеющим норму $\|x\|_S = c$, удовлетворяют равенству $x^T S^{-1} x = c^2$. Расстояние от точки x до центра кластера $\bar{x}$ определяется как

$$d_M(x,\bar{x}) = \sqrt{(x-\bar{x})^T S^{-1}(x-\bar{x})} \qquad (2)$$

Поскольку метрика Махаланобиса (2) не может быть применена в случае обращения в нуль хотя бы одного элемента главной диагонали матрицы S, то применяют ее модификации.

Заметим, что если матрица ковариации S в функции $d_M(x,y)$ является единичной, то расстояние Махаланобиса становится равным расстоянию Евклида.

Определение 4. Расстоянием Евклида-Махаланобиса (Euclidian-Mahalanobis Distance) между двумя точками $x = (x_1,...,x_p)^T$ и $y = (y_1,...,y_p)^T$ в пространстве R^p называется функция вида [8]:

$$d_{E-M}(x,y) = \sqrt{(x-y)^T (S+E)^{-1}(x-y)}, \quad (3)$$

где E - единичная матрица. Метрика (3) устраняет недостаток метрики Махаланобиса, поскольку элементы ее главной диагонали всегда больше нуля.

Определение 5. Полиномиальным расстоянием Махаланобиса (Polynomial Mahalanobis Distance) между двумя точками $x = (x_1,...,x_p)^T$ и $y = (y_1,...,y_p)^T$ в пространстве R^p называется функция вида [2]:

$$d_{M_{\delta^2}}(x,y) = \sqrt{(N-1)(x-y)^T U^T W_{\delta^2}^{-1} U(x-y)}, \quad (4)$$

где $A^T A = U^T W_{\delta^2}^{-1} U$ - есть результат сингулярного разложения симметрической матрицы $A^T A$, причем: $W_{\delta^2} = diag(\omega_1 + \delta^2,...,\omega_p + \delta^2)$, где δ^2 — малое положительное число

Будем различать два случая измерения расстояний. В первом случае имеем задачу определения расстояний между точками одного и того же класса (внутриклассовое расстояние). Второй случай охватывает чрезвычайно важные задачи определения расстояний между случайной точкой и классом, а также расстояние между классами.

2. Метрики Махаланобиса

В данном разделе приводятся доказательство того, что функции Махаланобиса являются метриками, что открывает путь к решению задач кластеризации. Доказательства взяты из диссертационной работы [6].

Лемма 1. Пусть $D_M(x,y)$ есть квадрат расстояния Махаланобиса $d_M(x,y)$ с положительно определенной матрицей $A \in R^{p\times p}$. Тогда существует несингулярная матрица $B \in R^{p\times p}$ такая, что для каждых $x, y, \in R^p$ имеет место $D_M(x,y) = \|Bx - By\|^2$.

Доказательство: Пусть A есть симметрическая положительно определенная матрица. Как известно из линейной алгебры, существует такая несингулярная матрица B, что $A = B^T B$. Откуда следует:

$$D_M(x,y) = (x-y)^T B^T B(x-y) = \\ = (Bx-By)^T (Bx-By) = \|Bx-By\|^2.$$

Теорема 1. Расстояние Махаланобиса $d_M(x,y)$ является метрикой [6].

Доказательство: Пусть $x, y, z \in R^p$. Пользуясь Леммой 1 получаем:

$$\sqrt{D_M(x,z)} = \|Bx-Bz\| = \|Bz-Bx\| = \sqrt{D_M(z,x)},$$
$$\sqrt{D_M(x,y)} = \|Bx-By\| \le \|Bx-Bz\| + \|Bz-By\| = \\ = \sqrt{D_M(x,z)} + \sqrt{D_M(z,y)}.$$

Следовательно, $d_M(x,y)=\sqrt{D_M(x,y)}$ отвечает требованию неравенства треугольника.

Представляет некоторый интерес **Лемма 2** [6].

Лемма 2. Для квадратов расстояний Махаланобиса D_M и для всех $x,y,z\in R^p$ имеет место: $D_M(x,y)\le 2D_M(x,z)+2D_M(z,y)$

Доказательство:

$$D_M(x,y)=\|Bx-By\|^2\le 2\|Bx-Bz\|^2+$$
$$+2\|Bz-By\|^2=2D_M(x,z)+2D_M(z,y).$$

Результаты работы [6] распространяются и на метрику Евклида-Махаланобиса. В работе [9] рассмотрены некоторые дополнительные особенности этой метрики, что позволяет аргументировано применять ее в задачах классификации. В работе [2] утверждается (без приведения доказательства) справедливость выполнения условий метрики для функции $d_{M_{\delta^2}}(x,y)$, что требует, на наш взгляд, проведения дополнительных исследований ее свойств.

До сих пор мы рассматривали расстояние между двумя точками одного класса, т.е. внутриклассовое расстояние. Для решения задач классификации и кластеризации необходимо иметь инструмент для измерения расстояний между классами и расстояний между точкой (представляющей единственный объект множества) и классом.

3. Измерение расстояний между классами

Общий подход к построению метрик на основе функций Махаланобиса требует знания матриц ковариаций всех классов. В методе дискриминантного анализа [10] для измерения расстояний между точкой и классами строится объединенная ковариационная матрица

$$S_0=\frac{1}{m_1+m_2-2}(S_1+S_2), \qquad (5)$$

где S_1 и S_2 - матрицы ковариаций для классов Ω_1 и Ω_2 соответственно, причем классы представлены матрицами значений признаков размерности $(m_1\times p)$ и $(m_2\times p)$ соответственно, Принцип объединения ковариационных матриц может быть использован для построения метрики Евклида-Махаланобиса.

Будем решать задачи определения расстояния между классами, а также расстояния между точкой (как частным случаем некоторого неизвестного класса) и классами.

Пусть заданы три класса *A,B,C* и их центры $\bar{x},\bar{y},\bar{z}$ в метрическом пространстве R^p. Выпишем условия для метрик, ориентированных на измерение межклассовых расстояний:

1) $d(A,B)\ge 0$
2) $d(A,B)=0\Rightarrow A=B$
3) $d(A,B)=0\Leftarrow A=B$
4) $d(A,B)=d(B,A)$
5) $d(A,B)\le d(A,C)+d(C,B)$

Построим матрицы ковариаций S_A,S_B,S_C. Далее удобно использовать матрицу, определенную не только при положительных, но и при нулевых значениях дисперсий. Например, обобщенная матрица для двух классов может быть задана, как сумма $S_{A,B}=S_A+S_B+E$, а для трех – соответственно $S_{A,B,C}=S_A+S_B+S_C+E$. Расстояние $d_M(A,B)$ удовлетворяет, так же, как и расстояние Евклида, условиям: $d_M(A,B)\ge 0$, $A=B\Rightarrow d_M(A,B)=0$ и является симметричным.

Утверждение 2. Расстояния Махаланобиса $d_M(r,A)$, $d_M(A,C)$ являются метриками, т.к.

$$d_M(r,A)\le d_M(A,B)+d_M(B,r),$$
$$d_M(A,B)\le d_M(A,C)+d_M(C,B).$$

Доказательство:

По аналогии с доказательством Теоремы 1 справедливо:

$$\sqrt{(r-\bar{y})^T S_{A,B}^{-1}(r-\bar{y})}\le\sqrt{(\bar{x}-\bar{y})^T S_{A,B}^{-1}(\bar{x}-\bar{y})}+$$
$$+\sqrt{(\bar{x}-\bar{y})^T S_{A,B}^{-1}(\bar{x}-\bar{y})},$$

$$\sqrt{(\bar{x}-\bar{y})^T S_{A,B,C}^{-1}(\bar{x}-\bar{y})}\le\sqrt{(\bar{x}-\bar{z})^T S_{A,B,C}^{-1}(\bar{x}-\bar{z})}+$$
$$+\sqrt{(\bar{z}-\bar{y})^T S_{A,B,C}^{-1}(\bar{z}-\bar{y})}.$$

Таким образом, объединение частных матрицы ковариаций обеспечивает выполнение свойств метрики для функций Махаланобиса (Евклида-Махаланобиса) в случае измерения межклассовых расстояний.

4. Задача начального размещения кластеров

Задача 1. Задача размещения кластеров (the maximal volume problem) формулируется следующим образом: как расположить точки на сфере радиуса r, чтобы сумма расстояний между ними была максимальна [3, 4, 11].

Решение задачи является предметом актуальных исследований и лежит в основе геометрического размещения кластеров. Покажем, что решением Задачи 1 будет равномерное расположение точек на окружности.

Теорема 2: (об оптимальном размещении кластеров на n-мерной сфере). Пусть $M_i \in R^m \ \ (i = 1_1, 2, ..., n)$ - точки, лежащие в m-шаре K(r) радиуса r, S – граница шара K(r), $d(M_i, M_j)$ - расстояние между точками M_i и M_j. Выражение $\sum_{1 \le i < j \le n} d(M_i, M_j)$ принимает наибольшее значение в случае, когда точки M_i **равномерно** расположены на границе m-шара – L. Доказательство теоремы можно разделить на две части.

Лемма 2.1: Пусть $M_i \in R^m \ \ (i = 1_1, 2, ..., n)$ - точки, лежащие в m-шаре K(r) радиуса r, S – граница шара K(r), $d(M_i, M_j)$ - расстояние между точками M_i и M_j. Численное значение выражения $\sum_{1 \le i < j \le n} d(M_i, M_j)$ полученного для точек расположенных в шаре не превосходит значения этого выражения для точек лежащих на сфере S.

Доказательство:

Пусть имеется некоторое размещение точек $M_i \ \ (i = 1_1, 2, ..., n)$ в m-шаре *K(r)*. Покажем, что любую точку $M_i \in K(r)$ можно сдвинуть на сферу *S* так, что величина $\sum_{1 \le i < j \le n} d(M_i, M_j)$ не уменьшится. Действительно, выберем произвольную точку M_i и проведем через нее хорду. Выполним параметризацию точки M_i на этой хорде, т.е. построим новую систему координат, в которой одна из осей координат, например, первая совпадает с хордой. Т.к. M_i будет двигаться вдоль хорды, то ее координаты – (t;0;…;0), а координаты точки $M_j = (a_1; a_2; ...; a_m)$. Тогда функция от параметра *t*, равная расстоянию от точки $M_j \in K(r)$ до точки M_i, лежащей на хорде имеет вид $d(M_i, M_j) = f(t) = \sqrt{(a_1 - t)^2 + a_2^{\ 2} + ... + a_m^{\ 2}}$. Функция $f(t)$ зависит только от переменной *t*, т.к. точка $M_j \in K(r)$ неподвижна. Функция $f(t)$ выпукла, т.к. ее вторая производная положительна.

Рассмотрим аналогичные функции для других точек. Сумма выпуклых функций $\sum_{j=1}^{n} d(M_i; M_j)$ является выпуклой. Наибольшем значение выпуклой функции находится на одном из концов отрезка. Т.е. сдвиг точек на сферу не уменьшает значения выражения $\sum_{1 \le i < j \le n} d(M_i, M_j)$.

Лемма 2.1 доказана. Таким образом доказана и первая часть Теоремы 2, на основании которой точки для Задачи 1 следует размещать на сфере.

Лемма 2.2. Выражение $\sum_{1 \le i < j \le n} d(M_i, M_j)$ принимает наибольшее значение в случае, когда точки M_i **равномерно** расположены на границе (поверхности) m-шара – L.

Покажем, что данная лемма справедлива для плоского случая.

Лемма 2.2* (об оптимальном размещении кластеров в круге)**:** Пусть $M_i \ \ (i = 1_1, 2, ..., n)$ - точки, лежащие в круге K(r) радиуса r, L – граница круга K(r), $d(M_i, M_j) = d_{i,j}$ - расстояние между точками M_i и M_j. Выражение $\sum_{1 \le i < j \le n} d(M_i, M_j)$ принимает наибольшее значение в случае, когда точки M_i **равномерно** расположены на границе круга – L.

Покажем теперь, что максимальной сумма расстояний *D* будет, если расположить точки на окружности равномерно. Распишем сумму *D* в виде:

$$D = \begin{cases} \sum_{p=1}^{m} \sum_{i=1}^{n} d_{i,i+p}, & n = 2m + 1; \\ \sum_{p=1}^{m-1} \sum_{i=1}^{n} d_{i,i+p} + \sum_{i=1}^{m} d_{i,i+m}, & n = 2m. \end{cases} \quad (6)$$

Введем систему координат с началом в центре нашего круга, тогда точки на окружности будут иметь координаты ($r\cdot cos\varphi$; $r\cdot sin\varphi$). Перенумеруем точки M_i так, чтобы точке M_k соответствовал угол φ_k, $k = 1,2,...,n$ и выполнялось соотношение:

$$0 = \varphi_1 \le \varphi_2 \le ... \le \varphi_k \le ... \le \varphi_n \le 2\pi .$$

Обозначим: $\varphi_{k+1} - \varphi_k = 2\beta_k$, $k = 1..n$, здесь $\varphi_{n+1} = 2\pi$.

Тогда длина хорды $M_k M_{k+1}$, $k = 1..n$, будет равна $d_{k,k+1} = 2r\sin\beta_k$.

При этом $\beta_1 + \beta_2 + ... + \beta_n = \pi$.

Заметим, что функция $f = \sin\beta$ вогнута на отрезке [0;2π]. Рассмотрим для начала сумму расстояний между соседними точками из (6):

$\sum_{k=1}^{n} d_{k,k+1} = \sum_{k=1}^{n} 2r\sin\beta_k$ По следствию из неравенства (Йенсена) [12, 13]: $\sum_{k=1}^{n}\sin\beta_k \le n\sin\frac{\beta_1 + ... + \beta_n}{n} = n\sin\frac{\pi}{n}$, т.е. эта сумма наибольшая, когда точки M_k расположены на окружности равномерно.

Теперь рассмотрим сумму расстояний между точками, расположенными через одну:

$\sum_{k=1}^{n} d_{k,k+1=2} = \sum_{k=1}^{n} 2r\sin(\beta_k + \beta_{k+1})$, здесь $\beta_{n+1} = \beta_1$.

Т.к. все n чисел вида $\beta_k + \beta_{k+1}$ принадлежат отрезку [0; π], то для этой суммы в силу неравенства Йенсена для вогнутой функции верна оценка сверху:

$$\sum_{k=1}^{n} d_{k,k+1=2} =$$

$$= \sum_{k=1}^{n} 2r\sin(\beta_k + \beta_{k+1}) \le 2nr\sin\frac{1}{n}\sum_{k=1}^{n}(\beta_k + \beta_{k+1}) =$$

$$= 2nr\sin\frac{2\pi}{n}. \qquad (7)$$

Равенство (7) соответствует равномерному размещению точек M_k на окружности L. Аналогичным образом позывается, что суммы расстояний точек через две, три и т.д. наибольшие, если точки M_k расположены равномерно на окружности. **Лемма 2.2* доказана**.

К сожалению, доказательство **Леммы 2.2.** для n-мерной сферы, вызывает затруднения, поэтому предлагается замена ее доказательства на решение интуитивно понятной вспомогательной **Задачи 2,** решение которой имеет самостоятельное и важное прикладное значение.

Задача 2. (задача о равномерном размещении точек на сфере). Необходимо **равномерно** распределить N точек по поверхности m-сферы радиусом R.

Прежде всего, необходимо утонить ***понятие равномерного размещения точек*** на m-сфере. Его в информационных источниках определяют по-разному.

Определение 6 (определение равномерности 1). Условие равномерного распределения на m-сфере выполняются, если точки представляют собой вершины правильных многогранников, описанных сферой. Правильный N-мерный многогранник — это выпуклый многогранник (т.е. расстояние между любыми ближайшими точками постоянно и равно ребру многогранника), но в таком случае количество решений задачи ограничено.

Рассмотрим случай трехмерной сферы.

Лемма 2.3. В случае трехмерной сферы и числе вершин N=4, условию равномерности 1 соответствует **правильный** тетраэдр.

Выберем систему координат так, чтобы первая вершина тетраэдра имела координаты $M_1 = (0;0;R)$. Выполним повороты вокруг оси OY на углы t_i, а затем последовательно на углы φ_i вокруг оси OZ для получения точек M_2, M_3 и M_4. При этом обобщенная матрица поворотов имеет вид $\begin{pmatrix} \cos t_i\cos\varphi_i & \cos t_i\sin\varphi_i & \sin t_i \\ -\sin\varphi_i & \cos\varphi_i & 0 \\ -\sin t_i\cos\varphi_i & -\sin t_i\sin\varphi_i & \cos t_i \end{pmatrix}$, а координаты полученных вершин тетраэдра: $M_i = (-R\sin t_i\cos\varphi_i; -R\sin t_i\sin\varphi_i; R\cos t_i)$, i = 2, 3, 4.

Тогда $d(M_1;M_i) =$

$$= R\sqrt{(\sin t_i\cos\varphi_i)^2 + (\sin t_i\sin\varphi_i)^2 + 1 - 2\cos t_i + (\sin t_i)^2} =$$

$$= 2R\sin\frac{t_i}{2},$$

а соответственно,

$$d_{1,2}+d_{1,3}+d_{1,4}=2R\sum_i \sin\frac{t_i}{2}\le 6R\sin\frac{t_2+t_3+t_4}{6}.$$

Т.к. это выражение зависит от среднего арифметического значения аргументов t_i, то естественно считать, что $t_2=t_3=t_4=t_0$ и $d_{1,2}+d_{1,3}+d_{1,4}=6R\sin(\frac{t_0}{2})$.

Теперь определим расстояния между точками M_i, $i=2,3,4$.

$$d(M_i;M_{i+1})=$$

$$=R\sqrt{(\sin t_i)^2(\cos\varphi_{i+1}-\cos\varphi_i)^2+(\sin t_i)^2(\sin\varphi_{i+1}-\sin\varphi_i)^2}$$

$$=2R\sin t_0\sin\frac{\varphi_{i+1}-\varphi_i}{2}$$

=> (в силу тех же соображений, что и в плоском случае)

$$2R\sin t_0\sum_i\sin\frac{\varphi_{i+1}-\varphi_i}{2}\le 6R\sin t_0\sin\frac{\pi}{3}=3\sqrt{3}R\sin t_0.$$

Следовательно, вся сумма расстояний будет равна

$$\sum_i d_{1,i}+\sum_i d_{i,i+1}=6R\sin(\frac{t_0}{2})+3\sqrt{3}R\sin t_0.$$

Определим, при каком значении t_0 эта сумма имеет наибольшее значение. После взятия производной и приравнивания ее нулю получим следующее равенство

$$\cos(\frac{t_0}{2})+\sqrt{3}\cos(t_0)=0 =>$$

$2\sqrt{3}\cos^2(\frac{t_0}{2})+\cos(\frac{t_0}{2})-\sqrt{3}=0$, откуда

$$\cos(\frac{t_0}{2})=\frac{-1+\sqrt{1+4\cdot 2\cdot 3}}{4\sqrt{3}}=\frac{1}{\sqrt{3}},\ \sin(\frac{t_0}{2})=\frac{2}{\sqrt{6}}.$$

Для правильного тетраэдра должны выполняться равенства

$a=\frac{4R}{\sqrt{6}}$, $r=\frac{\sqrt{6}}{12}a$, где a,R,r - соответственно ребро, радиусы описанной и вписанной окружности. Из геометрических построений с учетом найденного угла следует:

$$a=2R\sin(\frac{t_0}{2})=\frac{4}{\sqrt{6}}R,\quad r=R\cos(180-\alpha)=\frac{R}{3}=\frac{\sqrt{6}}{12}a,$$

т.е. тетраэдр правильный. Полученное решение составляет важный, но частный результат.

Решение **Задачи 2.** для больших размерностей на основе **Определения 6** вызывает большие затруднения. Поэтому целесообразно перейти к другим определениям равномерности.

Определение 7 (определение равномерности 2 по Томсону)**.** Расположим на сфере N одинаковых свободных зарядов. Будем считать распределение равномерным, если оно соответствует минимуму потенциальной энергии системы.

Для реализации этого способа используются специальные алгоритмы [14,15], основанные на принципе равновесия системы зарядов, предложенного физиком Томсоном.

5. Результаты экспериментальных исследований

Рассмотрим пример использования равномерного размещения кластеров в задаче коммивояжера. Задача имеет сложность порядка $n!$, где n - количество городов, и относится к классу *NP* — полных задач. Ее решение нейронной сетью Кохонена может быть основано на модели начального равномерного размещения m $(m>>n)$ точек – кластеров на окружности с нормированными координатами (x_i,y_i), $i=1,...,m$. Эти координаты интерпретируются как весовые коэффициенты (начальная настройка сети). Затем на вход сети подаются в случайном порядке входные вектора координат *n* городов, определяется в каждом случае кластер-победитель и производится коррекция его местоположения [11]. В результате ряда итераций прокладывается путь между городами. На Рис.1 отражены некоторые этапы решения задачи.

Места скопления городов (Рис.2а) могут создавать проблему, которая решается локальным перебором. После применения алгоритма получают достаточно приемлемый контур обхода, как это показано на Рис.2б.

В серии экспериментов, при случайном распределении точек–городов, было получено примерно равное количество лучших результатов, достигаемых метриками Евклида и Евклида-Махаланобиса. По результатам исследований разработано экспериментальное программное обеспечение для решения задач классификации и кластеризации текстовых и графических элементов на полутоновых снимках [16].

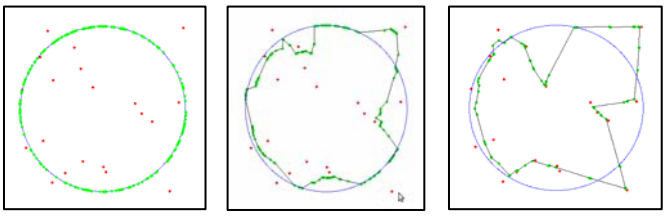
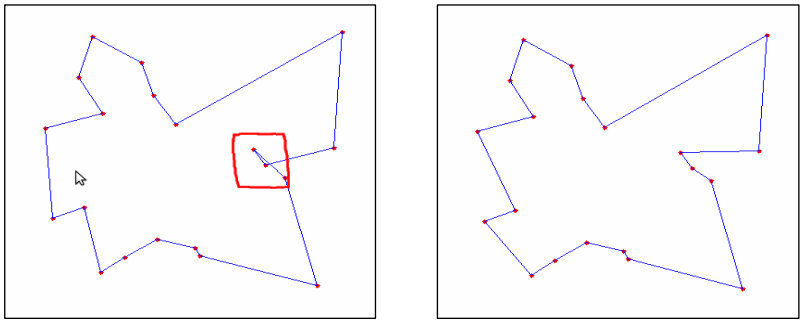


Рис. 1. Промежуточные этапы решения задачи коммивояжера



а) б)

Рис. 2. Результат решения задачи коммивояжера

Заключение

В настоящей работе даны определения расстояний и метрик, приведены доказательства того, что функции Махаланобиса, при определенных условиях, являются метриками, которыми можно пользоваться для решения задачи классификации. При этом преимущество в практическом использовании имеет метрика Евклида-Махаланобиса, которая обладает большей универсальностью. Для определения расстояний между точкой и классами, а также между классами необходимо проводить объединение матриц ковариаций. Показано, что решение **Задачи 1** может быть получено размещением точек на n-сфере. Принцип равномерного размещения кластеров в плоском случае экспериментально исследован при решении задачи коммивояжера. Решение задачи равномерного размещения точек на m-сфере наталкивается на большие сложности, которые преодолеваются численными методами.

Автор выражает искреннюю благодарность Алекберли М.М., Амелькину С.А. и Трушкову В.В. за помощь в работе.

Литература

1. Mahalanobis Distance. – http://classifion.sicyon.com/References/M_distance.pdf
2. Greg Grudic and Jane Mulligan, Outdoor Path Labeling Using Polynomial Mahalanobis Distance, Robotics: Science and Systems II, 2006. – http://www.roboticsproceedings.org/rss02/p20.pdf
3. Sloane N.J.A., Hardin R.H., Duff T.S., Conway J.H. Minimal-Energy Clusters. http://www.research.att.com/~njas/cluster/index.html
4. Hardin R.H., Sloane N.J.A., Smith W.D. Tables of Spherical Codes with Icosahedral Symmetry. http://www.research.att.com/~njas/icosahedral.codes/index.html
5. Многомерный статистический анализ в экономике: Учебное пособие для вузов./ Л.А. Сошникова, В.Н. Тамашевич, Г. Усбе, М. Шефер; Под ред. В.Н. Тамашевича. - М.: БНИТИ – Дана, 1999. – 598 с.
6. Ackerman M.R. Algorithms for the Bregman k-Median Problem. – A dissertation submitted to the Department of Computer Science University of Paderborn, 2009. – 220 pp.
7. Неравенства Коши. – http://works.tarefer.ru/50/100059/index.html#
8. Амелькин С.А., Захаров А.В., Хачумов В.М. Обобщенное расстояние Евклида-Махаланобиса и его свойства. – Информационные технологии и вычислительные системы, № 4, 2006, с. 40-44.
9. Хачумов М.В. О выборе метрики для решения задач классификации и кластеризации. – Материалы Первой всероссийской научной конференции с международным участием (SASM-2011) «Системный анализ и семиотическое моделирование» (Казань, 24-28 февраля 2011г.) – Казань: Издательство «Фэн» Академии наук РТ, 2011, с.255-260.

10. Методы дискриминантного анализа. – http://knowledge.allbest.ru/emodel/d-3c0b65625b3ac68a5d43a88421306c37.html
11. Атаманов В.В., Козачок М.А., Трушков В.В., Хачумов М.В. Выбор первоначального расположения кластеров в нейронной сети Кохонена. – Нейрокомпьютеры: разработка и применение, №1, 2009, с.73- 76.
12. Коровкин П. П. Неравенства. М., 1983. – 56 с.
13. Беккенбах Э., Беллман Р. Неравенства. – М.: Ком Книга, 2007. –276 с.
14. Андреев Н.Н., Юдин В.А. Экстремальные расположения точек на сфере. − Математическое просвещение (третья серия), 1997, вып.1, с.115-121. – http://www.etudes.ru/ru/mov/mov009/i2115125.pdf.
15. Андреев Н.Н. Минимальный дизайн 11 порядка на трехмерной сфере. – Математические заметки, Т. 67, Вып.4, апрель, 2000, с. 489-497.
16. Талалаев А.А., Тищенко И.П., Хачумов М.В. Выделение и кластеризация текстовых и графических элементов на полутоновых снимках. – Искусственный интеллект и принятие решений, № 3, 2008, с.72-84.

Хачумов Михаил Вячеславович. Аспирант РУДН. Окончил РУДН в 2009 году. Автор 12 печатных работ. Область научных интересов: искусственный интеллект, машинная графика, кластеризация. E-mail: khmike@inbox.ru, vmh48@mail.ru