

ЗАДАНИЕ МЕТРИКИ В ЗАДАЧАХ КЛАССИФИКАЦИИ ОБЪЕКТОВ РАЗЛИЧНОЙ ПРИРОДЫ

Рассматривается вопрос задания меры близости при классификации объектов различной природы. Проводится анализ эффективности использования евклидовой метрики в задачах классификации объектов различной природы. Рекомендуется для определения степени сходства объектов вместо коэффициентов сходства Рао, Хаммана, Дейка, Танимото использовать меру близости. Рассматривается пример практического применения меры близости для количественного определения степени сходства объектов.

1. Постановка задачи

Целью исследования является разработка меры близости для объектов, заданных числовым вектором. Классификации объектов различной природы, как правило, выполняется с помощью ЭВМ, что требует наличия четкого и достаточно простого алгоритма. В научных и прикладных сферах при классификации объектов или измерений используют коэффициенты сходства различных исследователей Рао, Хаммана, Дейка, Танимото [1,4]. Оперировать с коэффициентами сходства несложно, но эффективней применять меру близости [3,4]. Для решения конкретных задач классификации, чтобы определить, являются ли два объекта близкими между собой, необходимо дать количественное определение меры близости. Это достигается введением функции, измеряющей близость на множестве рассматриваемых объектов или измерений. Понятие близости является одним из основных в таких задачах и поэтому требует не интуитивного представления, а математически корректного.

2. Выбор меры близости классифицируемых объектов

Наиболее употребительной в настоящее время является евклидова мера, хотя она имеет существенный недостаток – не учитывает возможной неравномерности осей пространства. Обобщением евклидовой метрики является мера Махаланобиса, которая инвариантна относительно аффинных преобразований

$$d = \{(X_i - X_j)^T W^{-1} (X_i - X_j)\}^S, \quad (1)$$

где W^{-1} – матрица, обратная матрице рассеяния; X_i, X_j – числовые векторы измерений признаков, характеризующие соответственно i -й и j -й элементы множества объектов.

Выбор меры близости в значительной степени зависит от особенностей классифицируемых объектов. Так, для рассматриваемого в [2] множества элементов $X = \{X_i\}$, характеризующихся структурой отношений

$$X_i \cap X_j \neq \emptyset, X_i \notin X_j, |X_i| \neq |X_j|, i \neq j, \quad (2)$$

$$X_i = \{g_{ik}\}, g_{ik} \in \{0,1\}, i, j = \overline{1, n}, k = \overline{1, m},$$

в качестве меры близости использовалось выражение на основе коэффициента сходства Рао:

$$d_1 = 1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|}. \quad (3)$$

С точки зрения практических приложений для рассматриваемого выше множества элементов X , признаки которых являются двоичными переменными, могут оказаться полезными следующие метрики:

$$d_2 = 1 - \frac{|X_i \cap X_j|}{|X_i| + |X_j|}, \quad (4)$$

$$d_3 = 1 - \frac{2|X_i \cap X_j|}{|X_i| + |X_j|}. \quad (5)$$

Для общего случая, когда $g_{ip} \in \{0, 1, 2, \dots, k\}$, в качестве меры для группирования можно использовать выражение

$$d_{ij} = 1 - \frac{\sum_{p=1}^m \alpha_{ij}^p}{|X_i| + |X_j|}, \quad (6)$$

$$\text{где } \alpha_{ij}^p = \begin{cases} 0, & \text{если } g_{ip}g_{jp} = 0, \\ g_{ip} + g_{jp}, & \text{если } g_{ip}g_{jp} \neq 0. \end{cases}$$

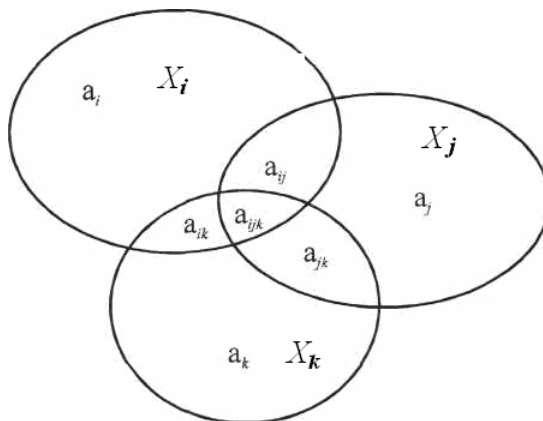
Чтобы выражение (6) использовалось в качестве меры близости, необходимо проверить выполнение аксиом Фреше.

Так как для любой пары X_i, X_j справедливо (2), то очевидно, что $0 \leq d_{ij} \leq 1$, $d_{ij} = d_{ji}$.

Необходимо проверить справедливость аксиомы треугольника

$$1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|} \leq 1 - \frac{|X_i \cap X_k|}{|X_i \cup X_k|} + 1 - \frac{|X_k \cap X_j|}{|X_k \cup X_j|}. \quad (7)$$

Для проверки выполнения аксиомы треугольника воспользуемся рисунком взаимных пересечений множества конструктивно-технологических признаков, характеризующих изделия X_i, X_j, X_k :



Обозначим взаимные пересечения множества признаков, характерные для объектов, представленных на рисунке:

$$a_i = |X_i \setminus [(X_i \cap X_j) \cup (X_i \cap X_k)]|, \quad (8)$$

$$a_j = |X_j \setminus [(X_j \cap X_i) \cup (X_j \cap X_k)]|, \quad (9)$$

$$a_k = |X_k \setminus [(X_k \cap X_i) \cup (X_k \cap X_j)]|, \quad (10)$$

где a_i, a_j, a_k – признаки, присущие соответственно только i -му, j -му, k -му объекту:

$$a_{ij} = |X_i \cap X_j \setminus [(X_i \cap X_j \cap X_k)]|, \quad (11)$$

здесь a_{ij} – признаки, одновременно присущие i -му и j -му объектам:

$$a_{ik} = |X_i \cap X_k \setminus [(X_i \cap X_j \cap X_k)]|, \quad (12)$$

a_{ik} – признаки, одновременно присущие i -му и k -му объектам:

$$a_{jk} = |X_j \cap X_k \setminus [(X_i \cap X_j \cap X_k)]|, \quad (13)$$

где a_{jk} – признаки, одновременно присущие j -му и k -му объектам:

$$a_{ijk} = |X_i \cap X_j \cap X_k|, \quad (14)$$

a_{ijk} – признаки, одновременно присущие i -му, j -му и k -му объектам:

$$X = |X_i \cup X_j \cup X_k| = a_i + a_j + a_k + a_{ij} + a_{ik} + a_{jk} + a_{ijk}. \quad (15)$$

Согласно взаимным пересечениям, из рисунка и с учетом выражений (8)-(15) неравенство (7) примет вид:

$$\frac{a_{ij} + a_{ijk}}{X - a_k} + \frac{a_{ik} + a_{ijk}}{X - a_j} - \frac{a_{jk} + a_{ijk}}{X - a_i} \leq 1. \quad (16)$$

3. Исследование выполнимости меры близости

Для проверки выполнимости неравенства (16) воспользуемся теоремой о необходимых условиях экстремума функции, заданной в виде неравенства [5].

Обозначим:

$$U = \frac{a_{ij} + a_{ijk}}{X - a_k} + \frac{a_{ik} + a_{ijk}}{X - a_j} - \frac{a_{jk} + a_{ijk}}{X - a_i}. \quad (17)$$

Составим функцию Лагранжа

$$F = -\lambda_0 \left(\frac{a_{ij} + a_{ijk}}{X - a_k} + \frac{a_{ik} + a_{ijk}}{X - a_j} - \frac{a_{jk} + a_{ijk}}{X - a_i} \right) - \lambda \left(\sum_e a_e^- - X \right) - \sum_e \lambda_e^- a_e^-, \quad \bar{e} = \{i, j, k, ij, ik, jk, ijk\}, \quad (18)$$

где λ_0, λ и λ_e^- – множители Лагранжа, согласно [5] не все равны нулю, при условии, что

$$\sum_e a_e^- - X = 0, \quad (19)$$

$$a_e^- \geq 0. \quad (20)$$

Так как ограничения (19) линейны, то из [3] следует, что

$$\lambda_0 = 1. \quad (21)$$

Тогда (18) будет иметь вид:

$$F = \frac{a_{jk} + a_{ijk}}{X - a_i} - \frac{a_{ij} + a_{ijk}}{X - a_k} - \frac{a_{ik} + a_{ijk}}{X - a_j} - \lambda \left(\sum_e (a_e^- - X) - \sum_e \lambda_e^- a_e^- \right). \quad (22)$$

Продифференцируем (22) по $a_{ij}, a_{ik}, a_{kj}, a_{ijk}, a_k, a_j, a_i, X$ и приравняем производные нулю:

$$-\frac{1}{X - a_k} - \lambda - \lambda_{ij} = 0, \quad (23)$$

$$-\frac{1}{X - a_j} - \lambda - \lambda_{ik} = 0, \quad (24)$$

$$\frac{1}{X - a_i} - \lambda - \lambda_{kj} = 0, \quad (25)$$

$$-\frac{a_{ij} + a_{ijk}}{(X - a_k)^2} - \lambda - \lambda_k = 0, \quad (26)$$

$$-\frac{1}{X - a_k} - \frac{1}{X - a_j} + \frac{1}{X - a_i} - \lambda - \lambda_{ijk} = 0, \quad (27)$$

$$-\frac{a_{ij}}{(X - a_k)^2} - \lambda - \lambda_k = 0, \quad (28)$$

$$-\frac{a_{ik} - a_{ijk}}{(X - a_j)^2} - \lambda - \lambda_j = 0, \quad (29)$$

$$-\frac{a_{kj} + a_{ijk}}{(X - a_i)^2} - \lambda - \lambda_i = 0, \quad (30)$$

$$-\frac{a_{ij} + a_{ijk}}{(X - a_k)^2} - \frac{a_{ik} + a_{ijk}}{(X - a_j)^2} + \frac{a_{kj} + a_{ijk}}{(X - a_i)^2} + \lambda = 0. \quad (31)$$

Из (23) видно, что

$$\frac{1}{X - a_k} > 0, \quad \lambda_{ij} \geq 0, \quad (32)$$

следовательно,

$$\lambda < 0. \quad (33)$$

Из (25) и (33) следует, что

$$\lambda_{kj} > 0, \quad a_{kj} = 0. \quad (34)$$

Из (30) и (34) следует, что

$$\lambda_i > 0, \quad a_i = 0. \quad (35)$$

Предположим, будто бы

$$a_j > 0, \quad \lambda_j = 0. \quad (36)$$

Тогда с учетом (21) выражение (27) примет вид

$$\lambda_{ij} - \frac{a_j}{X(X - a_j)} - \lambda_{ijk} = 0. \quad (37)$$

Так как

$$\lambda_{ijk} \geq 0, \quad (38)$$

то из (37) следует

$$\lambda_{ij} > 0, \quad a_{ijk} = 0. \quad (39)$$

Из (29) и (37) выводим

$$\lambda = -\frac{a_{ik} + a_{ijk}}{(X - a_j)^2}. \quad (40)$$

Из (31) в соответствии с (34), (39), (40) получаем

$$\frac{a_{ijk}}{(X - a_k)^2} - \frac{a_{ijk}}{X^2} = 0 \quad (41)$$

или

$$a_{ijk}a_k = 0, \quad (42)$$

если $a_{ijk} = 0$, то из (28) и (37) следует

$$\lambda_k = -\lambda > 0. \quad (43)$$

Тогда из (42) получается, что

$$a_k = 0. \quad (44)$$

Таким образом, имеем

$$a_k = a_i = 0, \quad a_{kj} = a_{ij} = 0. \quad (45)$$

При учете же (15) и (45) выражение (16) примет вид

$$\frac{a_{ij} + a_{ijk}}{X - a_k} + \frac{a_{ik} + a_{ijk}}{X - a_j} - \frac{a_{jk} + a_{ijk}}{X - a_i} = \frac{a_{ik} + a_{ijk}}{X - a_j} = 1. \quad (46)$$

Итак, функция (18) имеет безусловный максимум, равный единице. Исходя из (46) можно заключить, что выполняется неравенство (7). Следовательно, множество $X = \{X_i\}$, $i = \overline{1, n}$ с определенным выше расстоянием d_{ij} образует метрическое пространство.

Выводы. Предложены новые меры близости, отражающие естественные соотношения между сравниваемыми изделиями. Они характеризуются простой и ясной геометрической интерпретацией, а использование их обеспечивает исключительно четкое разделение. Эти метрики прошли апробацию в задачах планирования приборостроительного производства. Приведенные метрики могут быть использованы при анализе и синтезе структур сложных систем различной природы (технических, экономических, социологических и др.).

Список литературы: 1. Боннер Р.Е. Некоторые методы классификации. В кн.: Автоматический анализ сложных изображений. М.: Мир, 1969. 273 с. 2. Салыга В. И., Федоров А. А. Модель текущей специализации в задаче распределения квартальной программы // Электротехническая промышленность. 1977. Вып. 8 (454). С. 23-25 3. Федоров А.А., Федоров М. А. Об одной мере близости экономических объектов, описываемых числовым вектором // Вестник ХГПИ. 4. Федоров А. А. Об одной мере близости объектов в признаковом пространстве // АСУ. Харьков, ХАИ. 1979. Вып. 2. С. 125-127. 5. Гаибова М.А. Многокритериальная оптимизация инвестиционных проектов развития промышленных предприятий. Самара: ГУ, 2004. 137с. 6. Иваниенко В.В. Управление эффективностью использования ресурсов производства. Харьков: Изд. ХНЭУ, 2005. 368 с.

Поступила в редколлегию 16.06.2010

Федоров Андрей Алексеевич, канд. техн. наук, доцент кафедры организации производства и управления персоналом НТУ «ХПИ». Научные интересы: разработка моделей производственных процессов, проблемы классификации. Адрес: Украина, 61002, Харьков, ул. Фрунзе, 21, тел. 707-68-56.

Лопухин Юрий Владимирович, ст. преподаватель кафедры АПВТ ХНУРЭ. Научные интересы: проектирование программного обеспечения, автоматизации проектирования цифровых устройств. Адрес: Украина, 61166, Харьков, пр.Ленина, 14, тел. 70-21-326.

Скобликов Алексей Юрьевич, асп. НИПКИ «Молния» НТУ «ХПИ». Научные интересы: телекоммуникация. Адрес: Украина, 61002, Харьков, ул. Фрунзе, 21, тел. 707-68-56.