УДК 658.012.102

А.А. ФЕДОРОВ, канд. техн. наук, доц., НТУ «ХПИ»,
О.А. БУТЕНКО, канд. экон. наук, доц., МРИ

ЗАДАНИЕ МЕТРИКИ В ЗАДАЧАХ КЛАССИФИКАЦИИ

Рассмотрен вопрос задания меры близости при классификации объектов различной природы.

The article describes the task of establishment a measure of affinity at classification objects of the different nature.

Ключевые слова: классификация, мера близости

Введение. Классификации объектов различной природы, как правило, выполняется с помощью ЭВМ, что требует наличия четкого и достаточно простого алгоритма

Постановка проблемы. При решении конкретных задач классификации, для того чтобы можно было определить, являются ли два объекта близкими между собой, необходимо дать количественное определение меры близости (метрики). 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 [1, 5].

Методология. Наиболее употребительной в настоящее время является, евклидова метрика и метрика Минковского. Евклидова метрика обладает существенным недостатком - не учитывает возможной неравномерности осей пространства. Обобщением евклидовой метрики является мера Махаланобиса, которая инвариантна относительно аффинных преобразований

$$d = \{(X_i - X_j)^T W^{-1} (X_i - X_j)\}^{1/2}, \qquad (1)$$

где W^{-1} – матрица, обратная матрице рассеяния;

где X_i ,X_j – числовые векторы измерений признаков, характеризующие соответственно i –ый и j – ый элементы множества объектов.

В случае булевых признаков может быть удобной метрика Хемминга:

$$d_{ij} = \sum_{k=1}^{n} \left| x_{k_i} - x_{k_j} \right| \qquad (2)$$

На действительной плоскости в двумерном ортогональном пространстве $R^2 = R \times R$ для любых двух элементов $x = (x_1 x_2)$ и $y = (y_1 y_2)$ можно выбрать следующую метрику

$$d_{(xy)} = |x_1 - y_1| + |x_2 - y_2| \qquad (3)$$

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

$$X_i \cap X_j \neq \phi, \qquad X_i \not\subset X_j, \qquad |X_i| \neq |X_j|, \quad i \neq j$$
$$X_i = \{g_{ik}\}, \qquad g_{ik} \in \{0,1\}, \qquad i, j = \overline{1,n}, \qquad k = \overline{1,m} \qquad (4)$$

в качестве меры близости использовалось выражение на основе коэффициента сходства Рао [1]

$$d_1 = 1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|} \qquad (5)$$

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

$$d_2 = 1 - \frac{|X_i \cap X_j|}{|X_i| + |X_j|}\ , \qquad (6)$$

$$d_3 = 1 - \frac{2|X_i \cap X_j|}{|X_i| + |X_j|} \qquad (7)$$

Для булевых признаков могут оказаться полезными метрики на основе коэффициентов сходства Хаммана, Дейка или Танимото [1, , 9]:

$$d_x = 1 - \frac{|x_i \cap x_j| - |x_i \cup x_j \setminus x_i \cap x_j|}{|x_i \cup x_j|}; \qquad (8)$$

$$d_Д = 1 - \frac{2|x_i \cap x_j|}{2|x_i \cup x_j| + |x_i \cup x_j \setminus x_i \cap x_j|}; \qquad (9)$$

$$d_T = 1 - \frac{|x_i \cap x_j|}{|x_i| + |x_j| - |x_i \cup x_j \setminus x_i \cap x_j|} \qquad (10)$$

Для общего случая, когда $g_{ip} \in \{0,1,2,...,k\}$, (при работе с векторами, координатами которых являются произвольные вещественные числа) в качестве меры для группирования в [3] использовать следующее выражение:

$$d_{ij} = 1 - \frac{\sum_{p=1}^{m} \alpha_{ij}^{p}}{|X_i| + |X_j|} \qquad (11)$$

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

Используя теорему о необходимых условиях экстремума функции, заданной в виде неравенства, в [3, 4] показано выполнимость метрических свойств меры (6, 11).

Выводы. Меры близости (5, 11) использовались при решении задачи распределения производственной программы многономенклатурного цеха по плановым периодам различной длительности. Использование метрики позволило получать распределения, которые значительно снижали количество планов – учетных единиц в каждом периодов сравнении со случайными распределениями.

Меры близости (5 – 11) могут быть использованы при анализе и синтезе структур сложных систем различной природы (технических, экономических, социальных).

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», вып. 8 (454), 1977. с. 23-25.
3. *Федоров А. А., Федоров М. А.* Об одной мере близости экономических объектов, описываемых числовым вектором. Вестник ХГПУ, «Технический прогресс и эффективность производства» №21, 1997. **4**. *Федоров А. А.* Об одной мере близости объектов в признаковом пространстве. В сб. Автоматизированные системы управления, вып. 2, Харьков, ХАИ. 1979. **5**. Задачи классификации и их программное обеспечение. – М.: Наука, 1990. – 136с.

Надійшла до редколегії 28.10.10