Таким чином, в роботі наведено Y-подібну модель ЖЦ ПЗ, яка відображує паралельне проведення процесів формалізації вимог до проекту, проектування і кодування та відповідних їм фаз тестування. На основі цієї моделі запропонована класифікація тестів.

1. *ISO* 12207: 1995. – (ГОСТ Р – 1999). ИТ. Процессы жизненного цикла программных средств.
2. *Didkovska M.* Criteria for integration testing of component-based software // Электроника и связь. – 2004. – No 23. – С. 90–94.
3. *Майерс Г.* Искусство тестирования программ / Пер с англ. под ред. Б. А. Позина. – Москва: Финансы и статистика, 1982. – 172 с.

НТУ України "Київський політехнічний інститут"

Надійшло до редакції 25.10.2006

УДК 519.6

А. Г. Каграманян, В. П. Машталир, Е. В. Скляр, В. В. Шляхов

Метрические свойства разбиений множеств произвольной природы

(Представлено членом-корреспондентом НАН Украины Ю. Г. Стояном)

The interpretation of data content is closely connected with partition analysis. Different applications require different detailings of data partitions. For a system to be successful in a variety of problems, several partitions have to be ensured for cognitive-like techniques. A rational combination of low-level and high-level capabilities seems to be the most promising way to significantly improve the data understanding integrally. To reduce the gap between low-level features and high-level semantics in clustering, we propose, ground, and explore a new metric on partitions of an arbitrary measurable set.

В задачах распознавания образов факторизация информации в том или ином признаковом пространстве концептуально является одним из основных методов, лежащих в основе интерпретации данных. С одной стороны, может требоваться идентификация объектов, процессов или явлений с точностью до заданного или найденного в процессе анализа отношения эквивалентности. С другой, построение классов эквивалентности — часто суть и цель обработки данных для последующего этапа тематической трактовки отдельных регистрируемых представителей или даже фактор-множеств. В качестве типичного примера можно указать традиционную кластеризацию данных [1], особенно с целью дальнейшего компаративного распознавания [2]. Более детальным примером может служить сегментация изображений, т. е. построение разбиения поля зрения, удовлетворяющее условию принадлежности носителя "области интереса" одному классу эквивалентности [3]. В силу существенной неопределенности входных данных и возможности применения различных алгоритмов можно получать различные результаты, в частности, чрезмерную (объект расположен в нескольких факторах) или недостаточную (в классе эквивалентности находятся изображения фона) сегментацию.

Таким образом, достаточно очевидной и важной является задача получения инструментария для сравнения разбиений, отличия которых индуцированы либо вариативностью их получения, либо различием источников информации. Принципиальное значение сопоставление разбиений приобретает в задачах, когда характеристика факторов определяет вычислительную эффективность обработки и интерпретации данных в целом [4, 5]. В работе вводится функционал, являющийся метрикой на разбиениях произвольных измеримых множеств, и исследуются его свойства.

Постановка задачи. Пусть Ω — произвольное измеримое множество с мерой $\mu(\circ)$, которая может интерпретироваться как длина, площадь, объем, распределение масс, вероятности, мощность множества. Пусть Π_Ω — множество конечных (по количеству факторов) разбиений Ω, т. е. $\alpha \in \Pi_\Omega$, $\alpha = \{X_1, X_2, \ldots, X_n\}$, $X_i \subseteq \Omega$, $\mu(X_i) < \infty$, $i = \overline{1,n}$, $\Omega = \bigcup\limits_{i=1}^{n} X_i$, $\forall i, j \in \{1, 2, \ldots, n\}$: $i \neq j \Rightarrow X_i \bigcap X_j = \varnothing$. Требуется на $\Pi_\Omega \times \Pi_\Omega$ найти функционалы, являющиеся метриками, и изучить их свойства.

Метрики на разбиениях множеств. Следует подчеркнуть, что для установления мер сходства между разбиениями на концептуальном уровне необходимо учитывать степени различия (в частности, в виде симметрических разностей) и сходства (в частности, в виде пересечений) факторов. Доказана

Теорема 1. *Для произвольного измеримого множества Ω с мерой $\mu(\circ)$ и пары произвольных его разбиений $\alpha, \beta \in \Pi_\Omega$, $\alpha = \{X_1, X_2, \ldots, X_n\}$, $\beta = \{Y_1, Y_2, \ldots, Y_m\}$ функционал*

$$\rho(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mu(X_i \Delta Y_j) \mu(X_i \bigcap Y_j), \tag{1}$$

где $X_i \Delta Y_j = (X_i \setminus Y_j) \bigcup (Y_j \setminus X_i)$ — симметрическая разность, является метрикой.

При доказательстве теоремы 1 найдено равносильное представление метрики (1)

$$\rho(\alpha, \beta) = \sum_{i=1}^{n} [\mu(X_i)]^2 + \sum_{j=1}^{m} [\mu(Y_j)]^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{m} [\mu(X_i \bigcap Y_j)]^2. \tag{2}$$

Естественный интерес вызывает вопрос о связи метрики (1) с другими функционалами, оценивающими сходство-различие. Сформулируем гипотезу об общем виде класса метрик на произвольных разбиениях измеримых множеств.

Воспользуемся схемами геометрической вероятности. С одной стороны, разделим левую и правую часть (1) на $\mu(\Omega)^2$

$$\rho^*(\alpha, \beta) = \frac{\rho(\alpha, \beta)}{\mu(\Omega)^2} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\mu(X_i \Delta Y_j)}{\mu(\Omega)} \frac{\mu(X_i \bigcap Y_j)}{\mu(\Omega)} = \sum_{i=1}^{n} \sum_{j=1}^{m} p(X_i \Delta Y_j) p(X_i \bigcap Y_j). \tag{3}$$

Получаем: $p(X_i \Delta Y_j)$, $p(X_i \bigcap Y_j)$ — вероятности событий $X_i \Delta Y_j$ и $X_i \bigcap Y_j$ соответственно.

С другой стороны, напомним, что функционал, определяющий расстояние между двумя ансамблями данных Q и R, зависит от условной энтропии $H(Q, R)$ и взаимной энтропии $E(Q, R)$

$$\rho_H(Q, R) = H(Q, R) - E(Q, R). \tag{4}$$

Применительно к разбиениям метрика (4) конкретизируется следующим образом. Пусть $\Omega \subseteq \mathbb{R}^k$, тогда мы имеем два вероятностных распределения: $Q = (q_1, q_2, \ldots, q_n)$ и $R = (r_1, r_2, \ldots, r_m)$, где $q_k = \mu(X_k)/\mu(\Omega)$, $k = \overline{1,n}$, $r_l = \mu(Y_l)/\mu(\Omega)$, $l = \overline{1,m}$. Принимая во внимание, что $H(Q) = -\sum\limits_i q_i \ln q_i$, $H(Q,R) = -\sum\limits_i q_i \sum\limits_j r_j(q_i) \ln r_j(q_i)$ и $E(Q,R) = H(Q) + H(R) - H(Q,R)$, окончательно находим

$$\rho^*(\alpha,\beta) = \rho(P,Q) = -\sum_{k=1}^{m} p(X_k) \ln p(X_k) - \sum_{l=1}^{n} p(Y_l) \ln p(Y_l) + \\ + 2\sum_{k=1}^{m}\sum_{l=1}^{n} p(X_k \bigcap Y_l) \ln p(X_k \bigcap Y_l). \tag{5}$$

Перепишем (3) с учетом (2)

$$\rho(\alpha,\beta) = \sum_{i=1}^{n} [p(X_i)]^2 + \sum_{j=1}^{m} [p(Y_j)]^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{m} [p(X_i \bigcap Y_j)]^2. \tag{6}$$

Сравнивая (6) и (5), можно сделать предположение об общем виде метрики на разбиениях

$$\rho(\alpha,\beta) = \sum_{i=1}^{n} f(p(X_i)) + \sum_{j=1}^{m} f(p(Y_j)) - \sum_{i=1}^{n}\sum_{j=1}^{m} f(p(X_i \bigcap Y_j)). \tag{7}$$

Отметим, что в метрике (6) или, что равносильно, (1) фактически использована функция $f(x) = x^2$, а метрика (5) базируется на функции $f(x) = -x \ln x$. Вероятно, существуют и другие функции, индуцирующие метрики на произвольных разбиениях измеримых множеств. Вопрос о свойствах функций в (7), которые бы обеспечивали выполнение аксиом рефлексивности, симметричности и главное — неравенства треугольника, остается открытым.

Варианты интерпретации метрики (1) на разбиениях множеств. Возвращаясь к (3), проанализируем более общую вероятностную интерпретацию введенной метрики (1). Если рассматривать множество Ω как пространство элементарных исходов, являющееся элементом вероятностного пространства $\langle \Omega, \mathfrak{F}_\Omega, P \rangle$, где Ω — вообще говоря, произвольное множество; $\mathfrak{F}_\Omega$ — σ-алгебра его подмножеств, включающая в себя и Ω, и пустое множество, а P — вероятностная мера (или вероятность), относительно которой все элементы $\mathfrak{F}_\Omega$ измеримы, т. е. для любого $X \in \mathfrak{F}_\Omega$ существует число $0 \leqslant P(X) \leqslant 1$, удовлетворяющее аксиоматике вероятности по Колмогорову [6], то на исходном пространстве можно анализировать различные полные группы событий или наборы гипотез. Иными словами, функционал (1) в виде (6) является метрикой на парах различных наборов гипотез или полных группах событий.

Используя (3), можно ввести метрику и для сравнения дискретных случайных величин ξ' и ξ'', заданных на исходном конечном вероятностном пространстве

$$\overline{\rho}(\xi',\xi'') = \sum_{i=1}^{n} [p(\xi' = x_i)]^2 \left(1 - \sum_{j=1}^{m} [p(\xi'' = y_i)]^2\right) + \sum_{j=1}^{m} [p(\xi'' = y_i)]^2 \left(1 - \sum_{i=1}^{n} [p(\xi' = x_i)]^2\right). \tag{8}$$

Теорема 2. *Расстояние (в смысле функционала* (1)*) между двумя независимыми конечными равномерными распределениями с размерностями* n *и* m *имеет вид*

$$\overline{\rho}(\xi', \xi'') = \frac{1}{n} + \frac{1}{m} - \frac{2}{nm}.$$

Обратимся теперь к задачам, указанным во введении, которые привели к формализации постановки задачи. Прежде всего, важным моментом является то, что на множестве Ω, как правило, задана некоторая функция. Так, при сегментации изображений — это функция распределения яркостей, построение "правильных" разбиений носителя которой и отвечает задаче выделения из фона изображения объекта. В более общем случае, если в $\mathbb{R}^k$ задана некоторая неотрицательная функция $\varphi(x) \geqslant 0$, для которой $\int\limits_{\Omega} \varphi(x)dx < \infty$, $\Omega \subseteq \mathbb{R}^k$, то она индуцирует меру на $X \subseteq \Omega$ в виде интеграла $\mu(X) = \int\limits_{X} \varphi(x)dx$. Доказана

Теорема 3. *Для произвольной области* $\Omega \subseteq \mathbb{R}^k$ *и заданной на ней неотрицательной (за исключением множества меры ноль) интегрируемой функции* $\varphi(x)$ *имеет место неравенство*

$$\left(\int\limits_{\Omega} \varphi(x)dx\right)^2 \geqslant \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\int\limits_{X_i \cap Y_j} \varphi(x)dx\right)^2 + \sum_{i=1}^{n-1}\sum_{i'=i+1}^{n}\left(\int\limits_{X_i} \varphi(x)dx\right)\left(\int\limits_{X_{i'}} \varphi(x)dx\right) +$$
$$+ \sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}\left(\int\limits_{Y_j} \varphi(x)dx\right)\left(\int\limits_{Y_{j'}} \varphi(x)dx\right),$$

где $\alpha = \{X_i\}_{i=1}^{n}$, $\beta = \{Y_j\}_{j=1}^{m}$, $\alpha, \beta \in \Pi_\Omega$ — *два произвольных разбиения* Ω.

Доказанное неравенство позволяет формировать функцию невязки как разность левой и правой частей и формулировать задачу поиска, например, разбиения α при заданном β как задачу математического программирования, в которой ограничения индуцируются требованиями к результатам кластеризации или сегментации.

Из теоремы 3 непосредственно следуют три интерпретации: для гипотез, полной группы событий и дискретных случайных величин.

Следствие 1. *Если в вероятностном пространстве заданы два набора гипотез* $A_1 = \{S_i\}_{i=1}^{n}$ *и* $A_2 = \{T_j\}_{j=1}^{m}$, *то*

$$\sum_{i=1}^{n}\sum_{j=1}^{m}[p(S_i \bigcap T_j)]^2 + \sum_{i=1}^{n-1}\sum_{i'=i+1}^{n} p(S_i)p(S_{i'}) + \sum_{j=1}^{m-1}\sum_{j'=j+1}^{m} p(T_j)p(T_{j'}) \leqslant 1.$$

Следствие 2. *Если в некотором вероятностном пространстве произвольное событие* C *разбито на два набора попарно непересекающихся подсобытий* $\{X_i\}_{i=1}^{n}$, $\{Y_j\}_{j=1}^{m}$, *образующих полную группу, то*

$$\sum_{i=1}^{n}\sum_{j=1}^{m}[p(X_i \bigcap Y_j)]^2 + \sum_{i=1}^{n-1}\sum_{i'=i+1}^{n} p(X_i)p(X_{i'}) + \sum_{j=1}^{m-1}\sum_{j'=j+1}^{m} p(Y_j)p(Y_{j'}) \leqslant p[(C)]^2.$$

Следствие 3. *Если на одном вероятностном пространстве заданы две дискретные случайные величины ξ' и ξ'' с конечным числом значений, то*

$$\sum_{i=1}^{n}\sum_{j=1}^{m}[p(\xi'=x_i,\xi'=y_j)]^2+\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n}p(\xi'=x_i)p(\xi'=x_{i'})+$$

$$+\sum_{j=1}^{m-1}\sum_{j'=j+1}^{m}p(\xi''=y_j)p(p(\xi''=y_{j'}))\leqslant 1.$$

Введем на множестве разбиений Π_Ω функционал $\Phi(\alpha)=\sum\limits_{i=1}^{n}[\mu(X_i)]^2$, тогда справедлива

Теорема 4. *Для любых трех конечных разбиений $\alpha,\beta,\gamma\in\Pi_\Omega$ множества Ω выполняется*

$$\Phi(\gamma)\geqslant\Phi(\alpha\bigcap\gamma)+\Phi(\beta\bigcap\gamma)-\Phi(\alpha\bigcap\beta).$$

Здесь пересечение разбиений трактуется как совокупность традиционных пересечений отдельных факторов и соответствует более детальному представлению (измельчению) фактор-множеств.

Результаты и перспективы исследований. Введена и изучена метрика на произвольных разбиениях измеримых множеств. Показана ее связь с вероятностными интерпретациями расстояний для наборов гипотез и полных групп событий. Высказана гипотеза об общем виде класса метрик на разбиениях. Полезным направлением развития представляется конкретизация функционала (1) для сравнения разбиений, связанных отношением частичного порядка. Это позволит улучшить некоторые схемы иерархической кластеризации, основанные на ультраметриках. Особый интерес представляет нетривиальное (за счет вынесения пересечений в отдельные факторы) продолжение метрики (1) на конечные покрытия измеримых множеств, что создаст предпосылки для постановки оптимизационных задач факторизации произвольных измеримых множеств.

1. *Jain A. K., Dubes R. C.* Algorithms for clustering data. – Englewood Cliffs, New Jersey: Prentice Hall, 1988. – 320 p.
2. *Машталир В. П., Шляхов В. В.* Свойства мультиалгебраических систем в задачах компаративного распознавания // Кибернетика и систем. анализ. – 2003. – № 6. – С. 12–32.
3. *Форсайт Д., Понс Ж.* Компьютерное зрение. Современный подход. – Москва: Изд. дом "Вильямс", 2004. – 928 с.
4. *Kinoshenko D., Mashtalir V., Yegorova E.* Clustering method for fast content-based image retrieval // Computational Imaging and Vision / Ed M. A. Viergever. – Dordrecht: Springer. – 2006. – **32**. – P. 946–952.
5. *Kinoshenko D., Mashtalir V., Yegorova E., Vinarsky V.* Hierarchical partitions for content image retrieval from large-scale database // Machine Learning and Data Mining in Pattern Recognition / P. Perner., A. Imlya, eds. – Lecture Notes in Artificial Intelligence. – Berlin: Springer, 2005. – **3587**. – P. 445–455.
6. *Колмогоров А. Н.* Основные понятия теории вероятностей. – Москва: Наука, 1974. – 120 с.

Харьковский университет радиоэлектроники
Щецинский университет, Институт математики, Польша

Поступило в редакцию 10.11.2006