Эмпирическое построение решающих правил в стохастической морфологии*

А. В. ЗУБЮК
Московский государственный университет им. М. В. Ломоносова
e-mail: zubuk@cmpd2.phys.msu.su

УДК 519.68+681.513.7

Ключевые слова: морфологический анализ, стохастическая морфология, случайное множество, случайная форма, решающее правило, обучающая последовательность.

Аннотация

Задача классификации сигналов, имеющих случайную форму, рассмотрена как задача проверки статистических гипотез. Предложен метод эмпирического построения решающих правил в задачах проверки гипотез на основе обучающих последовательностей.

Abstract

A. V. Zubyuk, Empirical construction of decision procedures in stochastic morphology, Fundamentalnaya i prikladnaya matematika, vol. 15 (2009), no. 6, pp. 43—50.

Classification of signals with random forms is considered as statistical hypotheses test problem. A method for empirical construction of decision procedures in the statistical hypotheses test problems is proposed.

1. Методы морфологического анализа сигналов

В работах [3, 4] рассмотрены предложенные Ю. П. Пытьевым морфологические методы распознавания и идентификации сигналов, полученных при изменяющихся условиях регистрации. Такими сигналами могут быть, например, звуковые сигналы, электрические сигналы, изображения. Изменяющимися условиями регистрации могут быть освещение, при котором происходит оптическая съёмка сцены, спектральный диапазон, в котором работает регистрирующая аппаратура, и т. п. Методы морфологического анализа сформулированы в терминах инвариантов относительно изменений условий регистрации и поэтому позволяют распознавать и идентифицировать сигналы независимо от того, при каких условиях они получены.

Эти методы позволяют, например, узнавать какой-либо объект на оптических изображениях независимо от условий освещения, при которых производилась

*Работа выполнена при поддержке РФФИ, грант № 08-07-00133-а.

Фундаментальная и прикладная математика, 2009, том 15, № 6, с. 43—50.

съёмка. Вариации освещения при этом могут заключаться в изменении яркости освещения, угла падения света и т. п. Также методы морфологического анализа дают возможность совмещать изображения одной и той же сцены, полученные в разных спектральных диапазонах. Методами, изложенными в [3, 4], может быть решён целый ряд подобных задач анализа изображений.

Рассмотрим идею применения морфологических методов на примере задачи классификации изображений. Пусть $\mathcal{R}$ — пространство всех изображений и имеется N различных сцен, изображения которых, полученные при различных условиях наблюдения, предъявляются для классификации. Обозначим $V_1, V_2, \ldots, V_N$ множества в пространстве $\mathcal{R}$, каждое из которых определяет класс всех изображений, которые в рамках решаемой задачи считаются эквивалентными: изображения, принадлежащие к одному и тому же множеству V_i, являются изображениями одной и той же сцены, полученными при разных условиях регистрации. Такие множества эквивалентных изображений в морфологическом анализе принято называть *формами* изображений сцен. Пусть в этом случае требуется классифицировать предъявленное изображение g, т. е. узнать, изображением какой из N сцен оно является. Для этого надо определить, к какому из множеств $V_1, V_2, \ldots, V_N$ оно принадлежит, т. е., другими словами, какую форму оно имеет. Если же предъявляемое изображение g в процессе регистрации искажается случайным шумом, то для его классификации необходимо решить задачу проверки статистических гипотез о принадлежности g к одной из форм $V_1, V_2, \ldots, V_N$. Заметим, что изложенный подход применим не только при анализе изображений, но и при анализе любых оцифрованных сигналов.

Однако такие методы могут быть использованы лишь тогда, когда известны формы $V_1, V_2, \ldots, V_N$ регистрируемых сигналов, т. е. задана модель их формирования. Однако в ряде задач анализа сигналов такая модель существует, но не известна исследователю. При этом дана обучающая последовательность сигналов, представляющая собой наборы сигналов, про каждый из которых известно, какую из N форм он имеет. В таких случаях встаёт вопрос об эмпирическом построении модели формирования сигналов, которая затем позволит построить решающее правило классификации предъявляемых сигналов, либо об эмпирическом построении непосредственно правила классификации (без промежуточного этапа построения модели). В данной статье предложен один из возможных методов решения такой задачи.

2. Основные понятия стохастической морфологии. Задача проверки статистических гипотез

В ряде задач анализа сигналов форма сигнала может быть задана неточно. Например, рукописные изображения одной и той же буквы могут иметь различную геометрическую форму. Различными могут быть пропорции отдельных элементов буквы, наклон и т. п., причём варьирование указанных парамет-

ров может сделать изображение буквы плохо узнаваемым и даже похожим на изображение какой-либо другой буквы. В этом случае уместно было бы говорить, что такое изображение является изображением рассматриваемой буквы, но лишь с некоторой долей вероятности. Однако методы, описанные в [3, 4], не дают возможности учесть *вероятность* принадлежности некоторого изображения к классу изображений той или иной сцены, а позволяют сделать лишь бинарное заключение: принадлежит или нет.

Для описания вероятности принадлежности сигналов к некоторому классу вполне применим формализм случайных множеств. Основываясь на нём, дадим определение случайной формы сигнала как случайного множества в пространстве $\mathcal{R}$.

Определение 1. Случайной формой сигнала из пространства $\mathcal{R}$ назовём вероятностное пространство $(\mathcal{P}(\mathcal{R}), \mathcal{A}, \Pr)$, где $\mathcal{P}(\mathcal{R})$ — множество всех подмножеств пространства $\mathcal{R}$, $\mathcal{A}$ — некоторая σ-алгебра подмножеств множества $\mathcal{P}(\mathcal{R})$, а $\Pr$ — заданная на ней вероятность.

Рассмотрим задачу классификации предъявляемого сигнала в случае, когда формы сигналов случайны. Пусть дано N случайных форм

$$F_1 = (\mathcal{P}(\mathcal{R}), \mathcal{A}, \Pr_1),\ F_2 = (\mathcal{P}(\mathcal{R}), \mathcal{A}, \Pr_2), \ldots,\ F_N = (\mathcal{P}(\mathcal{R}), \mathcal{A}, \Pr_N),$$

где вероятности $\Pr_1, \ldots, \Pr_N$ заданы плотностями $\mathrm{pr}^{(1)}(\cdot), \ldots, \mathrm{pr}^{(N)}(\cdot)$ соответственно, и предъявляемый для идентификации сигнал ξ формируется по схеме

$$\xi = f + \nu, \tag{2.1}$$

где f — сигнал, принадлежащий одной из случайных форм $F_1, \ldots, F_N$, а ν — случайный элемент с нулевым математическим ожиданием и плотностью распределения $\mathrm{pr}_\nu(\cdot)$ (шум). Требуется по предъявленному сигналу ξ определить, какой из случайных форм принадлежит сигнал f.

Из (2.1) следует, что ξ — случайный элемент пространства $\mathcal{R}$ с плотностью распределения вида

$$\mathrm{pr}_\xi^{(t,\varphi)}(x) = \int\limits_{\omega \in \mathcal{P}(\mathcal{R})} \mathrm{pr}^{(t)}(\omega)\, \mathrm{pr}_\nu\big(x - \varphi(\omega)\big)\, \mathrm{d}\omega, \quad t = 1, \ldots, N, \tag{2.2}$$

где $\varphi(\cdot)\colon \mathcal{P}(\mathcal{R}) \to \mathcal{R}$ — функция, которая каждому подмножеству пространства $\mathcal{R}$ ставит в соответствие элемент из этого подмножества (т. е. $\varphi(\omega) \in \omega$, $\omega \subset \mathcal{R}$), а в остальном является произвольной. Различным функциям $\varphi(\cdot)$ отвечают различные плотности распределений в (2.2). Для каждого $t = 1, \ldots, N$ обозначим H_t множество всех распределений $\mathrm{pr}_\xi^{(t,\varphi)}(\cdot)$. Если $H_i \cap H_j = \varnothing$, $i \neq j$, то можно классифицировать предъявленный сигнал, решив задачу проверки статистических гипотез с альтернативами $H_1, H_2, \ldots, H_N$.

Минимаксный критерий (решающее правило) π для решения описанной выше задачи проверки статистических гипотез представляет собой набор N функций $\pi_i\colon \mathcal{R} \to [0, 1]$, $i = 1, \ldots, N$ (см. [1]). Этот критерий является решением

минимаксной задачи

$$\begin{cases} \max\limits_{i=1,\ldots,N} \alpha_i \sim \min\limits_{\{\pi_i\}}, \\ \sum\limits_{i=1}^{N} \pi_i(x) = 1, \quad x \in \mathcal{R}, \\ \pi_i(x) \geqslant 0, \quad x \in \mathcal{R}, \quad i = 1, \ldots, N, \end{cases} \tag{2.3}$$

где

$$\alpha_i \stackrel{\text{def}}{=} 1 - \min_{\mathrm{pr}_{\xi}^{(t,\varphi)}(\cdot) \in \mathcal{H}_i} \int\limits_{\mathcal{R}} \mathrm{pr}_{\xi}^{(t,\varphi)}(x) \pi_i(x)\, \mathrm{d}x, \quad i = 1, \ldots, N.$$

3. Решение минимаксной задачи проверки статистических гипотез

Рассмотрим для простоты случай, когда в минимаксной задаче (2.3) гипотезы H_i, $i = 1, \ldots, N$, простые. Это означает, что плотность распределения случайной величины ξ в (2.1) зависит только от номера альтернативы (обозначим эти распределения $\mathrm{pr}_{\xi}^{(t)}$, $t = 1, \ldots, N$) и критерий π является решением задачи

$$\begin{cases} \max\limits_{i=1,\ldots,N} \alpha_i \sim \min\limits_{\{\pi_i\}}, \\ \sum\limits_{i=1}^{N} \pi_i(x) = 1, \quad x \in \mathcal{R}, \\ \pi_i(x) \geqslant 0, \quad x \in \mathcal{R}, \quad i = 1, \ldots, N, \end{cases} \tag{3.1}$$

где

$$\alpha_i \stackrel{\text{def}}{=} 1 - \int\limits_{\mathcal{R}} \mathrm{pr}_{\xi}^{(i)}(x) \pi_i(x)\, \mathrm{d}x, \quad i = 1, \ldots, N.$$

Согласно [1] решением минимаксной задачи (3.1) будет рандомизированный критерий π, являющийся решением байесовской задачи

$$\sum_{i=1}^{N} q_i \alpha_i \sim \min_{\{\pi_i\}}, \tag{3.2}$$

где величины α_i определены в (3.1), а априорные вероятности q_i, $i = 1, \ldots, N$, таковы, что для критерия $\tilde{\pi}$, являющегося решением (3.2), выполнены равенства

$$\alpha_1 = \alpha_2 = \ldots = \alpha_N. \tag{3.3}$$

Как известно (см. [1]), решением задачи (3.2) при фиксированных $q_1, \ldots, q_N$ является критерий $\tilde{\pi}$, для которого

$$\begin{aligned}
&\tilde{\pi}_i(x) = 1, \quad \text{если} \quad q_i \operatorname{pr}_\xi^{(i)}(x) = \max_{j=1,\dots,N} q_j \operatorname{pr}_\xi^{(j)}(x), \\
&\tilde{\pi}_i(x) = 0, \quad \text{если} \quad q_i \operatorname{pr}_\xi^{(i)}(x) < \max_{j=1,\dots,N} q_j \operatorname{pr}_\xi^{(j)}(x), \\
&\sum_{i=1}^{N} \tilde{\pi}_i(x) = 1.
\end{aligned} \tag{3.4}$$

Основная проблема при нахождении решения задачи (3.2), (3.3) состоит в нахождении величинq_i, удовлетворяющих (3.3). Определив такие q_i, мы можем определить байесовский оптимальный критерий $\tilde{\pi}$, пользуясь (3.4). Этот критерий будет одновременно решением минимаксной задачи (3.1).

4. Эмпирическое построение решающих правил в задачах проверки статистических гипотез

Вероятностями $\Pr_1, \dots, \Pr_N$ (и их плотностями $\operatorname{pr}^{(1)}, \dots, \operatorname{pr}^{(N)}$), определяющими случайные формы $F_1, \dots, F_N$ из раздела 2, полностью определяется модель формирования предъявляемых сигналов в задачах стохастической морфологии, что даёт принципиальную возможность найти плотности $\operatorname{pr}_\xi^{(i)}$, $i = 1, \dots, N$, и решить задачу (3.2), (3.3).

Однако на практике модель формирования предъявляемых сигналов часто не определена, т. е. вероятности $\Pr_1, \dots, \Pr_N$ не заданы. Более того, вычисление интегралов в (2.2) и (3.1) может быть сопряжено с определёнными трудностями. Размерности пространств, по которым производится интегрирование, слишком велики, чтобы это интегрирование могло быть произведено численно. В задачах анализа изображений, например, размерность пространства $\mathcal{R}$ равна количеству пикселей на изображениях, т. е. может исчисляться сотнями тысяч и даже миллионами.

Независимо от причины, по которой невозможно точно вычислить распределения $\operatorname{pr}_\xi^{(i)}$, задача (3.2), (3.3) может быть решена приближённо путём эмпирического построения оптимального критерия $\tilde{\pi}$. Пусть имеется возможность получить выборки любого конечного объёма из распределений вероятностей $\Pr_i$. В случае когда $\Pr_i$ неизвестны, эти выборки могут быть получены из обучающей последовательности сигналов. В случае когда $\Pr_i$ известны, но невозможно произвести численное интегрирование в (2.2) и (3.1), эти выборки могут быть сгенерированы в результате численного эксперимента.

Требуется по полученным выборкам приближённо решить задачу (3.2), (3.3), которая, очевидно, эквивалентна задаче минимизации функционала

$$F \stackrel{\text{def}}{=} \max_{i,j=1,\dots,N} |\alpha_i - \alpha_j|$$

по параметрам q_i, где величины α_i вычисляются для критерия (3.4):

$$\max_{i,j=1,\dots,N} |\alpha_i - \alpha_j| \sim \min_{q_1,\dots,q_N}. \tag{4.1}$$

Решение задачи (4.1) может быть основано на следующеё теореме.

Теорема 1. *Пусть $\mathcal{R}$ — множество элементарных исходов, $\mathcal{A}(\mathcal{R})$ — σ-алгебра подмножеств множества $\mathcal{R}$ и на $\mathcal{A}(\mathcal{R})$ заданы абсолютно непрерывные вероятности $\Pr_1,\dots,\Pr_N$ с плотностями $\mathrm{pr}_1,\dots,\mathrm{pr}_N$. Обозначим Q множество всех наборов $q_1,\dots,q_N$ из N действительных чисел. Пусть вероятности $\Pr_1,\dots,\Pr_N$ таковы, что для любого $q \in Q$*

$$\Pr_t(\{x\colon q_i \,\mathrm{pr}_i(x) = q_j \,\mathrm{pr}_j(x),\ i,j = 1,\dots,N,\ i \neq j\}) = 0, \quad t = 1,\dots,N.$$

Пусть $\xi_1,\dots,\xi_N$ — независимые случайные величины, контролируемые вероятностями $\Pr_1,\dots,\Pr_N$ соответственно. Тогда для любого $q \in Q$, любого $\varepsilon > 0$ и любых натуральных чисел K и l с вероятностью не меньше чем $1 - 4N(KN+1)\exp(-2l\varepsilon^2)$ справедлива оценка

$$F^{K,l}_{\min}(q,\varepsilon) \leqslant F(q) \leqslant F^{K,l}_{\max}(q,\varepsilon) \tag{4.2}$$

и с вероятностью не меньше чем $1 - 4K\exp(-2l\varepsilon^2)$

$$\tilde{\mathcal{X}}^{K,l}_i(q,\varepsilon) \subset \mathcal{X}_i(q), \quad i = 1,\dots,N, \tag{4.3}$$

где

$$F(q) \stackrel{\text{def}}{=} \max_{i,j=1,\dots,N} |\alpha_i(q) - \alpha_j(q)|,$$

$$\alpha_i(q) \stackrel{\text{def}}{=} \Pr_i\bigl(\mathcal{R} \setminus \mathcal{X}_i(q)\bigr), \quad i = 1,\dots,N,$$

$$\mathcal{X}_i(q) \stackrel{\text{def}}{=} \{x \in \mathcal{R}\colon q_i \,\mathrm{pr}_i(x) > q_j \,\mathrm{pr}_j(x) \text{ для всех } j = 1,\dots,N,\ j \neq i\}.$$

В (4.2), (4.3) для любого наперёд заданного $\varepsilon > 0$, любого наперёд заданного разбиения $\mathcal{R}$ на K кластеров (подмножеств) $\mathcal{C}l^K_1,\dots,\mathcal{C}l^K_K$, $\mathcal{C}l^K_i \cap \mathcal{C}l^K_j = \varnothing$, $i,j = 1,\dots,K$, $i \neq j$, $\bigcup\limits_{i=1}^{K} \mathcal{C}l^K_i = \mathcal{R}$, и любого объёма l выборок из распределений вероятностей $\Pr_1,\dots,\Pr_N$ величины $F^{K,l}_{\min}(q,\varepsilon)$ и $F^{K,l}_{\max}(q,\varepsilon)$ и множества $\tilde{\mathcal{X}}^{K,l}_i(q,\varepsilon)$ определяются по следующему алгоритму.

1. *Для каждой случайной величины $\xi_1,\dots,\xi_N$ проводятся l независимых испытаний. Результатом являются N выборок объёма l из распределений вероятностей $\Pr_1,\dots,\Pr_N$.*
2. *По полученным выборкам определяются частоты $\nu^l_i(\mathcal{C}l^K_j)$ кластеров $\mathcal{C}l^K_1,\dots,\mathcal{C}l^K_K$ (частота ν^l_i соответствует выборке из распределения вероятности $\Pr_i$) и строятся множества*

$$\tilde{\mathcal{X}}^{K,l}_i(q,\varepsilon) \stackrel{\text{def}}{=} \bigcup_{j \in J^l_i(q,\varepsilon)} \mathcal{C}l^K_j, \tag{4.4}$$

где

$$J_i^l(q,\varepsilon) \stackrel{\text{def}}{=} \big\{ j\colon q_i(\nu_i^l(\mathcal{C}l_j^K) - \varepsilon) > q_a(\nu_a^l(\mathcal{C}l_j^K) + \varepsilon) \\ \text{для всех } a = 1,\dots,N,\ a \neq i \big\}, \quad i = 1,\dots,N.$$

3. *Вычисляются*

$$\tilde{\alpha}_i^{K,l}(q,\varepsilon) \stackrel{\text{def}}{=} \nu_i^l\Bigg(\bigcup_{j=1,\dots,N,\ j\neq i} \tilde{\mathcal{X}}_j^{K,l}(q,\varepsilon) \Bigg), \quad i = 1,\dots,N.$$

4. *Строится числовое множество*

$$A^{K,l}(q,\varepsilon) \stackrel{\text{def}}{=} \bigcup_{i,j=1,\dots,N} \big[|\tilde{\alpha}_i^{K,l}(q,\varepsilon) - \tilde{\alpha}_j^{K,l}(q,\varepsilon)| - 2\varepsilon,\ |\tilde{\alpha}_i^{K,l}(q,\varepsilon) - \tilde{\alpha}_j^{K,l}(q,\varepsilon)| + 2\varepsilon \big].$$

5. *Вычисляются искомые величины*

$$F_{\max}^{K,l}(q,\varepsilon) = \sup A^{K,l}(q,\varepsilon),$$
$$F_{\min}^{K,l}(q,\varepsilon) = \sup\big([\inf A^{K,l}(q,\varepsilon),\ \sup A^{K,l}(q,\varepsilon)] \setminus A^{K,l}(q,\varepsilon) \big).$$

Путём выбора параметров K, l и ε с помощью оценки (4.2) на основе обучающей выборки в задаче классификации сигналов можно сколь угодно точно и с любой заданной вероятностью оценить целевой функционал в задаче (4.1). Таким образом, с помощью теоремы 1 задача (4.1) превращается в задачу стохастического программирования, в которой целевой функционал известен с погрешностью. Методы решения таких задач подробно рассмотрены в [2].

Результатом решения указанной задачи стохастического программирования будет набор величин $q_1,\dots,q_N$, на которых достигается минимум в (4.1). При этом согласно (4.3) на объединении множеств $\tilde{\mathcal{X}}_i^{K,l}(q,\varepsilon)$, $i = 1,\dots,N$, критерий (3.4) с вероятностью не меньше чем $1 - 4K\exp(-2l\varepsilon^2)$ совпадает с критерием

$$\tilde{\pi}_i^{K,l,\varepsilon}(x) = \begin{cases} 1, & \text{если } x \in \tilde{\mathcal{X}}_i^{K,l}(q,\varepsilon), \\ 0, & \text{если } x \in \bigcup\limits_{j=1,\dots,N,j\neq i} \tilde{\mathcal{X}}_j^{K,l}(q,\varepsilon). \end{cases} \tag{4.5}$$

При выполнении условий теоремы 1 вероятности множества $\mathcal{R} \setminus \bigcup\limits_{i=1}^{N} \tilde{\mathcal{X}}_i^{K,l}(q,\varepsilon)$, определяемые плотностями распределений $\mathrm{pr}_\xi^{(i)}$, $i = 1,\dots,N$, могут быть сделаны сколь угодно малыми (за счёт уменьшения самого этого множества) выбором параметров K, l и ε и способа разбиения R на кластеры $\mathcal{C}l_1,\dots,\mathcal{C}l_K$. Поэтому почти для любого сигнала g из $\mathcal{R}$ параметры K, l и ε и кластеры $\mathcal{C}l_1,\dots,\mathcal{C}l_K$ могут быть выбраны так, чтобы $g \in \bigcup\limits_{i=1}^{N} \tilde{\mathcal{X}}_j^{K,l}(q,\varepsilon)$. Таким образом, теорема 1 позволяет эмпирически построить решающее правило в задаче классификации сигналов почти для любого сигнала из $\mathcal{R}$. Этим решающим правилом является критерий (4.5). При этом вероятность того, что критерий (4.5) совпадёт

с критерием (3.4) на множестве $\bigcup_{i=1}^{N} \tilde{\mathcal{X}}_j^{K,l}(q,\varepsilon)$ может быть сделана сколь угодно близкой к единице за счёт использования достаточно большой обучающей последовательности.

Предложенный метод эмпирического построения решающего правила, основанного на обучающей выборке, позволяет решать задачи классификации сигналов, имеющих случайную форму. Случайная форма определена как случайное множество сигналов. В качестве примеров приведём следующие задачи: классическая задача распознавания рукописных букв, задача классификации аэрокосмических изображений, задача классификации томографических снимков в медицинской диагностике, классификация электрокардиограмм.

Литература

[1] Боровков А. А. Математическая статистика. Оценка параметров. Проверка гипотез. — М.: Наука, 1984.

[2] Ермольев Ю. М. Методы стохастического программирования. — М.: Наука, 1976.

[3] Пытьев Ю. П. Морфологические понятия в задачах анализа изображений // ДАН СССР. — 1975. — Т. 224, № 6. — С. 1283—1286.

[4] Pyt'ev Yu. P. Morphological image analysis // Pattern Recognition and Image Analysis. — 1993. — Vol. 3, no. 1. — P. 19—28.