Е.А. Чернецова, А.Д. Шишкин

КЛАССИФИКАЦИЯ ПРОСТРАНСТВЕННО-РАСПРЕДЕЛЕННЫХ ОБЪЕКТОВ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ ПОСЛЕДОВАТЕЛЬНОЙ ВЫБОРКИ ИЗМЕРЕНИЙ

E.A. Chernetsova, A.D. Shishkin

SPACE-DISTRIBUTED OBJECT CLASSIFICATION VIA SUCCESSIVE SAMPLE MEASUREMENT USING

Рассматривается проблема классификации аномалий на морской поверхности. Предлагается для решения данной проблемы использовать модифицированный последовательный алгоритм Вальда. Приводится решающее правило и результаты моделирования предложенного алгоритма.

Ключевые слова: морские аномалии, классификация объектов, решающее правило, охрана окружающей среды.

A problem of sea surface anomalies classification is considered. For its decision it is suggested to use the modified sequential Vald algorithm. A decision rule and the algorithm modeling results are given.

Key words: marine anomalies, classification of objects, the decision rule, environmental protection.

Введение

Актуальнейшим вопросом гидроэкологии является борьба с загрязнением Мирового океана методами дистанционного обнаружения и оконтуривания пятен нефти и нефтепродуктов на поверхности акваторий. Для обнаружения нефтяных и иных аномалий на морской поверхности разрабатываются радиофизические методы, основанные на принципе различия контрастности оптических, тепловых и радиоактивных свойств гидроповерхности «чистой» воды и загрязненной нефтью и нефтепродуктами.

Главной проблемой обнаружения аномалий является то, что пространственные распределения их на морской поверхности выражены неявно. Поэтому необходимо сначала осуществить сегментацию области с целью выявления местоположения аномалии [1], определить ее параметры, и использовать их в качестве массива входных данных на этапе решения задачи классификации. Пример такой сегментации приведен на рис. 1.

В выделенном сегменте с помощью сканирования узким лучем по двум координатам измеряются в элементах разрешения интенсивности парциальных сигналов, которые могут позволить либо выявить контуры неоднородностей, либо определить области с одинаковой интенсивностью.

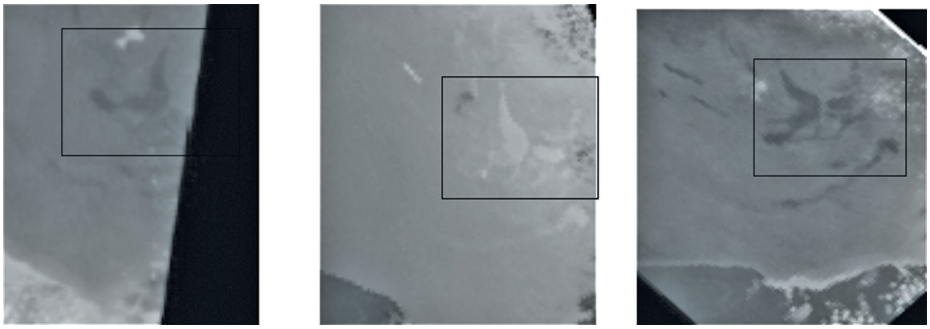

Рис. 1. Изображения нефтяного загрязнения, полученного спутниковым спектрорадиометром при разных углах визирования [1].

Так как априорная информация о распределении неоднородностей, как правило, недоступна, то для выявления контраста необходимо производить измерения, как от участка чистой воды, так и от аномалий. В случае радиолокационного обнаружения аномалия («темный объект») и фон имеют различные значения интенсивности отраженного сигнала. Выявления неоднородности может рассматриваться как задача классификации объектов, имеющих различную интенсивность, на основе определенных требований. В простейшем случае она понимается как разделение объектов на два класса. Считаем, что классификация заключается в вынесении решения: есть аномалия или нет ее. Простейший подход, по-видимому, заключается в сравнении измеряемых признаков с эталонами. Недостаток этого метода- отсутствие подходящего эталона.

Более совершенный подход классификации основывается на множестве отобранных замеров, производимых измерителем признаков. Такие замеры будем называть признаками. Они предполагаются инвариантными или малочувствительными к шумам, обладающими небольшой избыточностью. Считаем, что проблема измерения и отбора признаков решена на этапе обнаружения. На вход классификатора поступают две совокупности признаков: одна от эталона, другая от измерителя.

Классификацию по эталонам, можно рассматривать как частный случай, при чем набор признаков, измеренных от сегмента с чистой поверхностью, можно рассматривать как эталон. Допустим, что у каждого входного образа измеряется N признаков. Каждое множество из N признаков можно рассматривать как вектор X, называемый вектором признаков (замеров), или как точку в N-мерном пространстве признаков Ω_x. Задача классификации заключается в распределении всех возможных векторов или точек в пространстве признаков по соответствующим классам образов. Это можно трактовать как разбиение пространства признаков на взаимно непересекающиеся области, каждая из которых соответствует некоторому классу образов.

Теоретический анализ

Математически задача классификации может быть сформулирована с помощью разделяющей функции[2]. Пусть ω_1, ω_2 обозначают два возможных класса образов,

подлежащих классификации, и пусть $X = |x_1, x_2, ..., x_n|$ есть вектор замеров признаков, где x_i представляет собой i-й замер. Тогда разделяющая функция $D_j(X), j = 1,2$, относящаяся к классу образов ω_j, такова, что если входной образ, представленный вектором признаков X, принадлежит классу ω_j, то величина $D_j(X)$ должна быть наибольшей. Пусть $X \subset \omega_j$ обозначает, что вектор признаков X входного образа принадлежит классу ω_j. Тогда можно записать, что для всех $X \subset \omega_j$.

$$D_j(X) > D_i(X), \quad i \neq j. \tag{1}$$

Таким образом, в пространстве признаков Ω_x граница разбиений, называемая решающей границей, между областями, относящимися соответственно к классу ω_1 и классу ω_2 выражается следующим уравнением:

$$D_1(X) - D_2(X) = 0. \tag{2}$$

В (1) и (2) предполагалось, что измерения признаков дают детерминированные величины $X = |x_1, x_2, ..., x_n|$ и разделение на два класса может быть осуществлено с помощью одного порогового логического элемента. Если области признаков линейно разделимы гиперплоскостью, то можно получить вполне правильную классификацию. Однако в реальных случаях измеренные признаки образов одного класса могут претерпевать большие изменения, и, кроме того, нельзя пренебрегать помехами, возникающими при измерениях. При этих условиях можно рассматривать $|x_1, x_2, ..., x_n|$ как случайные величины, где x_i — результат измерения i-го признака в условиях помех.

Если предположить, что для каждого класса образов ω_j, известны многомерная (N-мерная) функция плотности вероятности $p(X/\omega_j)$ вектора признаков X и плотность априорной вероятности $P(\omega_j)$ появления ω_j, то решение задачи классификации можно сформулировать на основе байесовского правила минимизации потерь или среднего риска. Средний риск по существу является также вероятностью ложной классификации. Байесовское решающее правило дает:

$$X \subset \omega_1,$$

если

$$P(\omega_1)\, p(X/\omega_1) \geq P(\omega_2)\, p(X/\omega_2). \tag{3}$$

Или, переходя к отношению правдоподобия между классами ω_1 и ω_2:

$$\lambda = \frac{p(X/\omega_1)}{p(X/\omega_2)}, \tag{4}$$

где λ — пороговое значение, принимаемое как отношение:

$$\lambda \geq \frac{P(\omega_2)}{P(\omega_1)}. \tag{5}$$

Классификатор, осуществляющий байесово решающее правило, называют байесовым классификатором. Разделяющая граница классов для ω_1 и ω_2 байесового правила представляется следующим образом:

$$P(\omega_1)\,p(X/\omega_1) - P(\omega_2)\,p(X/\omega_2) = 0$$

или

$$\log\frac{P(\omega_1)\,p(X/\omega_1)}{P(\omega_2)\,p(X/\omega_2)} = 0. \qquad (6)$$

Если $p(X/\omega_i)$ есть функция многомерного гауссова распределения со средним вектором M_i и ковариационной матрицей K_i и если $K_1 = K_2$, то (6) приводится к каноническому виду:

$$X^T K^{-1}(M_1 - M_2) - \tfrac{1}{2}(M_1 + M_2)^T K^{-1}(M_1 - M_2) + \log\frac{P(\omega_1)}{P(\omega_2)} = 0. \qquad (7)$$

Следует отметить, что, байесово решающее правило при функции потерь (0,1) является также решающим правилом безусловного максимума правдоподобия. Более того, решение по (условному) максимуму правдоподобия можно рассматривать как байесово решающее правило при равных априорных вероятностях, т.е. при $P(\omega_j) = 1/2$, и $j = 1,2$.

В статистических системах классификации, описанных ранее, все N признаков «наблюдались» классификатором одновременно за один шаг. Это называют решающей процедурой с фиксированным объемом выборки. При этом фактически не учитывалась стоимость измерений признаков. Очевидно, что недостаточное число измерений признаков не позволит получить удовлетворительные результаты классификации. С другой стороны, практически нецелесообразно измерять чрезмерно большое число признаков. Если должна учитываться стоимость выполнения измерений признаков или если признаки входных образов в классификатор поступают последовательно, возможно применить последовательную решающую процедуру к этому классу задач классификации объектов [1, 2].

Такой подход особенно уместен при высокой стоимости выполнения измерений признаков. Существует связь между количеством информации, получаемым при измерении признаков и стоимостью выполнения этих измерений. Рациональное соотношение между принятием решения и числом измеряемых признаков можно получить, осуществляя измерение признаков последовательно и заканчивая этот последовательный процесс (принимая решение) когда достигнута достаточная или необходимая вероятность классификации.

Так как измерения признаков производятся последовательно, то здесь важен порядок, в котором измеряются признаки. Естественно, что признаки должны быть расположены в таком порядке, чтобы измерения дали окончательное решение возможно раньше. Задача упорядочения признаков является специальной задачей в системах последовательной классификации.

В случае двух классов объектов, можно применить последовательный критерий отношения вероятностей Вальда (*п. к. о. в.*) [2]. Этот критерий применяется для решения о выборе между двумя простыми гипотезами. Рассмотрим принцип *п. к. о. в.*

Предположим, что случайная величина z обладает функцией плотности $p(z/\theta)$, где θ — измеряемый параметр. Формулируем две гипотезы:

$$H_1: \quad \theta = \theta_1,$$

$$H_2: \quad \theta = \theta_2.$$

Построенный критерий решает дело в пользу θ_1 или θ_2 на основе наблюдений $Z = z_1, z_2, \ldots$ Допустим, что если гипотеза H_1 истинна, мы хотим получить решение, в ее пользу с вероятностью не меньшей $(1 - e_{21})$, а если истинна H_2, то решение в ее пользу должно иметь вероятность, не меньшую $(1 - e_{12})$.

Для постоянного объема выборки оптимальное решающее правило будет аналогично правилу (6) и (7).

$$\lambda_n = \prod_{i=1}^{n} \frac{p(z_i/H_1)}{p(z_i/H_2)} = \frac{p_n(Z/H_1)}{p_n(Z/H_2)}. \tag{8}$$

Этот критерий решает принять или отвергнуть гипотезу H_1, когда λ_n соответственно меньше или больше некоторой постоянной. Значение этой постоянной может быть выбрано так, чтобы критерий давал нужную величину e_{21}, и в принципе можно выбрать n так, чтобы критерий имел мощность $(1 - e_{12})$. Заметим, что e_{21} и e_{12} являются так называемыми «ошибкой первого рода» и «ошибкой второго рода».

Последовательный критерий отношения вероятностей Вальда аналогичен этому и обладает аналогичными оптимальными свойствами. Процедура испытания следующая: наблюдения производится до тех пор, пока выполняется условие:

$$B < \lambda_n < A. \tag{9}$$

Прекращаем наблюдения и принимаем решение в пользу гипотезы H_1, как только будет выполнено условие:

$$\lambda_n \geq A. \tag{10}$$

Прекращаем наблюдения и принимаем решение в пользу гипотезы H_2, как только будет выполнено условие:

$$\lambda_n \leq B. \tag{11}$$

Постоянные A и B называются соответственно верхним и нижним порогами. Они могут быть выбраны так, чтобы приблизительно получить заданные вероятности ошибок e_{21} и e_{12}. Как доказано [2] постоянные A и B определяются по формулам:

$$A \le \frac{1-e_{21}}{e_{12}}, \quad B \ge \frac{e_{21}}{1-e_{12}}. \tag{12}$$

При доказательстве (12) принималось, что признаки независимы.

Важным является также определить среднее количество измерений. Из [2] следует, что среднее число наблюдений, когда H_1 истинно, равно:

$$E_1(n) = \frac{(1-e_{21})\log A + e_{21}\log B}{E_1(z)}. \tag{13}$$

Аналогично для H_2:

$$E_2(n) = \frac{e_{12}\log A + (1-e_{12})\log B}{E_2(z)}. \tag{14}$$

Выражения (13) и (14) при знаках равенства представляют собой решающие границы, которые разбивают пространство признаков на три области:

- область, относящуюся к ω_1;
- область, относящуюся к ω_2;
- область неопределенности (или нулевую область), заключенную между двумя границами.

Область неопределенности соответствует условиям, когда не может быть принято окончательное решение. Естественно, что при последовательном процессе классификации решающие границы изменяются с числом замеров признаков n.

В предположении, например, что x_1, x_2, ... — независимые замеры признаков с одномерной гауссовой функцией плотности $p(x_j/\omega i)\, j = 1, 2, ..., i = 1, 2$ со средним значением m_i и дисперсией σ^2 последовательная решающая процедура Вальда имеет вид [2]:

Если $$\sum_{i=1}^{n} x_i \ge \frac{\sigma^2}{m_1 - m_2}\log A + \tfrac{1}{2}(m_1 - m_2), \quad \text{то} \quad x_1 \subset \omega_1, \tag{15}$$

если $$\sum_{i=1}^{n} x_i \le \frac{\sigma^2}{m_1 - m_2}\log B + \tfrac{1}{2}(m_1 - m_2), \quad \text{то} \quad x_1 \subset \omega_2, \tag{16}$$

и если $$\frac{\sigma^2}{m_1 - m_2}\log B + \tfrac{1}{2}(m_1 - m_2) < \sum_{i=1}^{n} x_i < \frac{\sigma^2}{m_1 - m_2}\log A + \tfrac{1}{2}(m_1 - m_2), \tag{17}$$

то берется $n + 1$ измерение.

Недостатком описанной процедуры является то, что параметры m_i и σ считаются известными, постоянными и не меняются от шага к шагу. В случае, если они неизвестны, то применение (*п. к. о. в.*) правила становится проблематичным. Кроме того, неизвестными являются и априорные вероятности $P(\omega_1)$ и $P(\omega_2)$.

Представляет интерес снять ряд ограничений, а именно: 1) правило решения должно приниматься при последовательной выборке; 2) мощность правила должна

быть независимой от неизвестной дисперсии и монотонно зависимой от среднего значения выборок.

Модификация и моделирование алгоритма

Предлагается внести в (*п. к. о. в.*) правило следующую модификацию.

Пусть $x_1, x_2, \ldots$ и $y_1, y_2, \ldots$ две последовательные выборки из нормальных распределений $N(\xi, \sigma_1^2)$ и $N(\eta, \sigma_2^2)$соответственно. Сначала по ограниченным выборкам размера m_0 вычислим оценки математического ожидания и дисперсии наблюдаемых процессов:

$$m_1 = \overline{x_0} = \frac{1}{m_9}\sum_{i=1}^{m_0} x_i, \qquad m_2 = \overline{y_9} = \frac{1}{m_9}\sum_{i=1}^{m_0} y_i,$$

$$S_1^2 = \frac{1}{m_0}\left[\sum_{i=1}^{m_0}\left(x_i - \overline{x_0}\right)^2\right], \quad S_2^2 = \frac{1}{m_0}\left[\sum_{i=1}^{m_0}\left(y_i - \overline{y_0}\right)^2\right].$$

Формируем из $x_1, x_2, \ldots$ и $y_1, y_2, \ldots$ объединенную выборку $z = z_1, z_2, z_3, \ldots$ объема $n = m_0$, где $z_i = x_i - y_i$, $i = 1, 2, \ldots$. Полученные оценки подставляем в правило (17):

$$\frac{s_2^2}{m_1 - m_2}\log B + \tfrac{1}{2}\left(m_1 - m_2\right) < \sum_{i=1}^{n} z_i < \frac{s_1^2}{m_1 - m_2}\log A + \tfrac{1}{2}\left(m_1 - m_2\right).$$

при выполнении которого, следует брать очередное $n_0 + 1$ измерение. Измерения проводятся до тех пор, пока не будет принято решение в пользу гипотезы H_1 или H_2.

Для проверки применимости и работоспособности рассмотренного алгоритма был проведено моделирование в пакете Matlab для двух сигналов, представленных своими гауссовскими плотностями распределения вероятностей. Результаты классификации этих сигналов по критерию Вальда приведены в таблице. Моделирование проводилось для сигналов, представленных своими нормальными ПРВ с параметрами $m_1 = 1$; $\sigma_1 = 2$; $m_2 = 1{,}1$; $\sigma_2 = 3$.

1	2	3	4	5	6	7
N	20	200	2000	20000	2000	2000
Σ_1	0,5986	0,5997	0,5997	0,5997	0,5997	0,5997
Σ_2	0,5685	0,5796	0,5796	0,5796	0,5796	0,5796
n	-0,0012	-0,0168	-0,168	-0,168	-1,1054	0,5815
v	0,002	0,0152	0,152	0,152	1,8549	0,5925
A	0,9	0,9	0,9	0,9	0,5	1,44
B	1,1	1,1	1,1	1,1	3,2	1,45
ПР	Сигналы принадлежат к одному классу	Сигналы принадлежат к одному классу	Сигналы принадлежат к одному классу	Сигналы принадлежат к одному классу	Неопределенность	Сигналы принадлежат к разным классам

В таблице введены следующие обозначения:
N — количество отсчетов ПРВ сигналов, взятых для моделирования;
Σ_1, Σ_2 — численное значение суммы элементов для первого и второго сигналов соответственно;
n, v — численные значения нижнего и верхнего порогов критерия соответственно;
A, B — параметры модели, соответствующие областям ошибок I и II рода;
ПР — принимаемое решение о классификации сигналов на основе моделирования.

Результаты моделирования позволяют сделать следующие выводы:

1. Увеличение количества отсчетов, взятых для моделирования, расширяют границы принятия решения для отнесения сигналов к одному и тому же или к разным классам, поэтому в случае получения неопределенности по критерию Вальда необходимо увеличивать количество измерений.
2. Расширение границ принятия решений в критерии Вальда, зависящее от числа измерений имеет свой предел (см., например, данные в столбцах 4 и 5 табл. 1), по достижению которого критерий Вальда зависит от параметров A и B, численные значения которых зависят от априорных знаний о вероятностях ошибок I и II рода
3. В случае большой похожести ПРВ сравниваемых сигналов, принадлежащих к разным классам (как, например, в табл. 1), с помощью критерия Вальда можно принять правильное решение (см. столбец 7 табл. 1), если обладать априорными знаниями о вероятностях ошибок I и II рода, выражающихся в правильном подборе параметров A и B модели.

Выводы

Если между параметрами ПРВ классифицируемых сигналов наблюдаются большие различия (например, $m_1 = 1$; $\sigma_1 = 2$; $m_2 = 10$; $\sigma_2 = 5$), то моделирование критерия Вальда позволяет принять решение о принадлежности исследуемых сигналов к разным классам при небольшом размере выборки: $N \leqslant 5$. Увеличение размера выборки согласно закону больших числе вызывает стирание различий между наблюдаемыми ПРВ сигналов и, соответственно, ошибочное решение о принадлежности наблюдаемых сигналов к одному и тому же классу.

Наиболее интересны результаты эксперимента по классификации сигналов, имеющих близкие по значению параметры ПРВ, хотя сами сигналы могут принадлежать как к одному и тому же, так и к разным классам.

Литература

1. *Pelyushenko S.A.* “The Use of Microwave Radiometer Scanning System for Detecting and Identification of Oil Spills”, in Proceedings of the Fourth Thematic Conference on Remote Sensing for Marine and Coastal Environments, Environmental Research Institute of Michigan, Ann Arbor, Michigan, 1997, vol. I, pp. 381–385.
2. *Фу К.* Последовательные методы в распознавания образов и обучении машин. — М.: изд. «Наука», 1971. — 256 с.
3. *Прокофьев В.Н., Шишкин А.Д.* О последовательной классификации нормальных совокупностей с неизвестными дисперсиями. // Радиотехника и электроника, 1974, т. 19, № 7, с. 2035–2038.