Исследование возможности классификации инфразвуковых сигналов методами проверки статистических гипотез

А. И. Чуличков[1,a], Н. Д. Цыбульская[1,b], С. Н. Куличков[2]

[1] *Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра компьютерных методов физики. Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2.*
[2] *Институт физики атмосферы имени А. М. Обухова РАН. Россия, 119017, Москва, Пыжевский пер., д 3.*
E-mail: [a] *achulichkov@gmail.com,* [b] *sandratsy@list.ru*

Статья поступила 09.12.2011, подписана в печать 15.12.2011

Исследуется возможность классификации сигналов методами проверки статистических гипотез. На основании анализа характерных особенностей сигналов, принадлежащих каждому классу, осуществлялось эмпирическое построение формы класса. Предложен механизм определения отделимости сигналов каждого класса, а также уровня критерия, определяющего критическую область. На основании исследования выведен алгоритм классификации. Эффективность предложенной методики проверена на задаче разделимости инфразвуковых сигналов, регистрируемых в атмосфере.

Ключевые слова: классификация сигналов, проверка статистических гипотез, морфологический анализ, эмпирическое построение формы.

УДК: 519.95. PACS: 02.50.Le.

Введение

В настоящей работе рассматривается вариант объединения двух подходов к обработке сигналов в применении к задаче классификации. Первый подход — морфологический анализ изображений [1, 2] — позволяет выделить характерные особенности сигналов и эмпирически построить форму для каждого класса. Под формой в методах морфологического анализа понимается информация, общая для элементов данного класса и не зависящая от условий регистрации. Например, в случае, когда неизвестен коэффициент усиления сигнала, форма должна быть инвариантной к изменениям амплитуды сигнала.

Второй подход связан с методами проверки статистических гипотез [3]. Он определяет возможность разделения классов и является основой алгоритма классификации.

Классический подход к задачам проверки статистических гипотез связан с именами К. Пирсона и Ю. Неймана [3] и состоит в задании решающего правила, позволяющего по наблюдению ξ принимать или отвергать гипотезу. Решающее правило выбирается таким образом, чтобы как можно реже ошибаться, принимая (неверную) гипотезу, допуская при этом, что в определенном проценте случаев мы будем ошибаться, отвергая (верную) гипотезу.

Задача классификации в терминах анализа статистических гипотез

Рассмотрим задачу разделения векторов на два класса. Предполагается, что элементы каждого класса являются случайными векторами из евклидова пространства R_n с нулевым математическим ожиданием и корреляционной матрицей V для первого класса и W для второго. Для решения задачи классификации воспользуемся нерандомезированным критерием, разбивающим пространство R_n на две области — область принятия гипотезы S и дополнение к ней, называемое критической областью $\overline{S}$. Если реализация случайного вектора попадает в область S, то она относится к первому классу, иначе — ко второму.

Область S будем строить из следующих соображений. Предположим, что верна гипотеза. Рассмотрим базис $\{\boldsymbol{e}_j,\ j=1,\ldots,n\}$ Карунена–Лоэва, составленный из собственных векторов матрицы V, соответствующий собственным значениям $\sigma_j^2,\ j=1,\ldots,n$, упорядоченных так, что $\sigma_1^2 \geqslant \sigma_2^2 \geqslant \ldots \geqslant \sigma_n^2$, тогда случайный вектор $\boldsymbol{\xi} \in R_n$ с нулевым математическим ожиданием и ковариационной матрицей V запишется в виде $\boldsymbol{\xi} = \sum\limits_{j=1}^{n} \alpha_j \boldsymbol{e}_j$, где коэффициенты разложения α_j — некоррелированные случайные величины с нулевым математическим ожиданием и дисперсией, равной $\sigma_j^2,\ j=1,\ldots,n$ [2]. После преобразования с помощью матрицы $V^{-1/2}$ получим вектор $V^{-1/2}\boldsymbol{\xi} = \sum\limits_{j=1}^{n} \frac{\alpha_j}{\sigma_j} \boldsymbol{e}_j$, коэффициенты разложения которого имеют единичную дисперсию, а квадрат его нормы $t(\boldsymbol{\xi}) = \left\| V^{-1/2}\boldsymbol{\xi} \right\|^2 \equiv \left(\boldsymbol{\xi}, V^{-1}\boldsymbol{\xi}\right)$ имеет математическое ожидание, равное размерности n пространства R_n. Тогда на основании неравенства Чебышева для любого числа $\varepsilon > 0$ можно записать $P(t(\boldsymbol{\xi}) \geqslant \varepsilon) \leqslant n/\varepsilon$.

Последнее соотношение используем для характеристики согласия реализации $\boldsymbol{x}$ случайного вектора $\boldsymbol{\xi} \in R_n$ с гипотезой. Подставив $\varepsilon = t(\boldsymbol{x})$, получим $P(t(\boldsymbol{\xi}) \geqslant t(\boldsymbol{x})) \leqslant n/t(\boldsymbol{x})$, что можно интерпретировать следующим образом: чем больше значение $t(\boldsymbol{x})$, полученное для реализации $\boldsymbol{x}$, тем меньше вероятность того, что при верной гипотезе появится значение $t(\boldsymbol{\xi})$, превосходящее $t(\boldsymbol{x})$. Значение $\alpha_V(\boldsymbol{x}) = n/t(\boldsymbol{x})$ является верхней гранью вероятности получить реализацию $\boldsymbol{\xi}$, согласующуюся с гипотезой не лучше, чем $\boldsymbol{x}$. Случайная величина $\alpha_V(\boldsymbol{x})$ носит название надежности гипотезы и используется как характеристика согласия реализации $\boldsymbol{x}$ с гипотезой [4].

Рассуждая аналогично, получим, что согласие реализации вектора $\boldsymbol{x}$ случайного вектора $\boldsymbol{\xi} \in R_n$ с альтернативой дается величиной $\alpha_W(\boldsymbol{x}) = n/(\boldsymbol{x}, W^{-1}\boldsymbol{x})$.

Так как ошибки первого и второго рода приводят к разным потерям, будем считать, что вектор $\boldsymbol{\xi}$ по реализации $\boldsymbol{x}$ относится к гипотезе, если разность $\alpha_V(\boldsymbol{x}) - \alpha_w(\boldsymbol{x}) \geqslant c$, где пороговое значение является параметром задачи, регулирующим соотношение между ошибками первого и второго рода. Сделав соответствующие преобразования, получим, что область S принятия гипотезы определится следующим соотношением:

$$S = \left\{\boldsymbol{x} \in R_n\colon\ (\boldsymbol{x}, V^{-1}\boldsymbol{x}) - (\boldsymbol{x}, W^{-1}\boldsymbol{x}) \leqslant c_\alpha\right\}. \qquad (1)$$

Эмпирическое построение формы классов

Для построения формы все сигналы, принадлежащие данному классу, разбивались на участки по методу «гусеницы» [5]. Полученные вектора рассматривались как реализации случайных векторов размерности n. Математические ожидания случайных векторов полагались равными нулю, а сами выборочные векторы нормировались.

По полученной выборке векторов первого класса строилась выборочная ковариационная матрица V. Число выборочных векторов равно $(N-n)kL$ (L — число сигналов выбранного класса, k — количество датчиков, регистрировавших сигнал, N — число отсчетов).

Векторы, аналогичным образом полученные для второго класса, рассматривались как выборочные значения случайного вектора, распределенного согласно альтернативе, и по ним строилась выборочная ковариационная матрица W.

Проверка разделимости классов и определение критических уровней

Для проверки разделимости для каждого сигнала i-го класса вычислялась функция $d_i(c_\alpha) = \left\{\text{число векторов } \boldsymbol{x}\colon\ (\boldsymbol{x}, V^{-1}\boldsymbol{x}) - (\boldsymbol{x}, W^{-1}\boldsymbol{x}) \leqslant c_\alpha\right\}$. Далее вычислялись оценки вероятности верного принятия гипотезы $P_1(c_\alpha) = d_1(c_\alpha)/N_1$ (N_1 — число векторов первого класса) и оценка вероятности неверного принятия альтернативы $P_2(c_\alpha) = d_2(c_\alpha)N_2$ (N_2 — число векторов второго класса).

По полученным данным можно, во-первых, определить возможность разделения классов, а во-вторых, указать пороги c_α, задающие критерий (1).

Алгоритм классификации

Окончательная классификация при выбранных порогах проводилась по следующему алгоритму.

1. Для классифицируемого сигнала методом «гусеница» строились $(N-n)kL$ выборочных векторов.

2. Каждый выборочный вектор классифицировался на основании критерия (1).

3. Считалось, что сигнал можно уверенно отнести к классу с номером i, если сумма числа участков сигнала, отнесенных к этому классу, деленная на число векторов класса, превышала некоторое пороговое значение h.

Эмпирическое построение модели классов сигналов

Эффективность метода проверялась на задаче классификации инфразвуковых сигналов [4]. Библиотека SigLib, содержащая эти данные, состоит из 57 сигналов, разделенных на 5 классов: взрыв (класс № 1, ExplosionTest), горные обвалы (класс № 2, MAW), микробаромы (класс № 3, Microbarom), вулканическая деятельность (класс № 4, VOL) и полярные сияния (класс № 5, AIW). Регистрация производилась 3–4 датчиками.

Для построения ковариационных матриц каждый сигнал разбивался на участки, кратные периоду ($n = 600$ отсчетов). Анализ полученных сигналов показал, что их математические ожидания близки к нулю, а матрицы ковариаций близки к «теплицевым». Таким образом, случайные векторы заданного класса можно рассматривать как реализации стационарных случайных процессов.

Из анализа разделимости на пять классов была выявлена хорошая классификация сигналов на два множества. В первое множество вошли сигналы 1-го и 4-го класса; во второе — сигналы 2-го, 3-го и 5-го класса.

Полученные графики $P_1(c_\alpha)$ и $P_2(c_\alpha)$ для случая разделения на два множества приведены на рис. 1. Величина уровня c_α выбрана равной 1500.

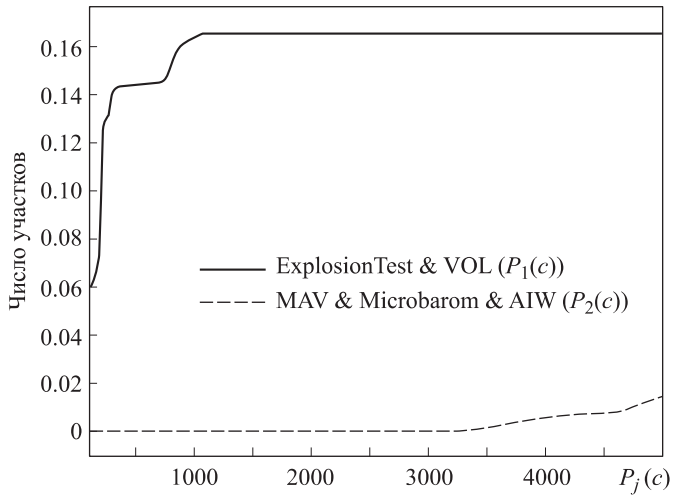

Рис. 1. Графики числа участков, для которых критерий меньше либо равен величине уровня c_α

Результаты классификации сигналов на два множества представлены на гистограмме рис. 2. По оси абсцисс обозначены классы, к которым относятся сигналы, оотенками серого выделены множества, к которым относятся сигналы. Считалось, что, если высота столбика превосходит величину $h = 0.8$, сигнал относится к множеству № 1, иначе — к множеству № 2.

Заключение

Для решения задачи классификации был использован подход, связанный с теорией статистического принятия гипотез [3]. Эмпирическая модель класса строилась методом «гусеницы», при этом сигнал разбивался на участки одинаковой длительности (600 отсчетов) [5].

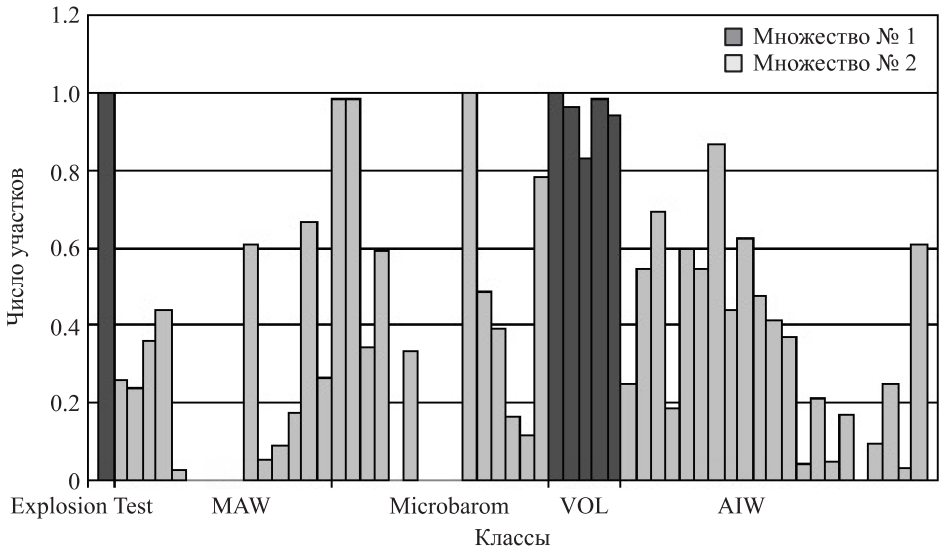

Рис. 2. Гистограмма разделения сигналов на два множества при $c_\alpha = 1500$

В результате было обнаружено достаточно хорошее различение между двумя объединенными множествами сигналов. На контрольной выборке при отнесении сигналов к этим двум множеством все сигналы из первого множества (взрывы и вулканическая деятельность) были классифицированы верно. Из второго множества (микробаромы, горные обвалы и полярные сияния) 3 из 51 сигнала были ошибочно отнесены к сигналам первого множества. Полученные результаты свидетельствует о хорошем качестве алгоритма.

Работа выполнена при финансовой поддержке РФФИ (гранты 11-07-00338-а, 11-05-00890, ГК № 70/ГФ/Н-11, ГК № 14.740.11.0203).

Список литературы

1. *Пытьев Ю.П., Чуличков А.И.* Методы морфологического анализа изображений. М., 2010.
2. *Чуличков А.И., Демин Д.С., Куличков С.Н.* Морфологический анализ инфразвуковых сигналов в акустике. М., 2010.
3. *Леман Э.* Проверка статистических гипотез. М., 1979.
4. *Пытьев Ю.П.* Методы математического моделирования измерительно-вычислительных систем. М., 2004.
5. *Голяндина Н.Э.* Метод «Гусеница»-SSA: анализ временных рядов. СПб., 2004.

Analysis of classification possibility infrasound signals from different sources based on correlation ability

A. I. Chulichkov [1,a], **N. D. Tsybulskaya** [1,b], **S. N. Kulichkov** [2]

[1] *Department of Computational Methods in Physics, Faculty of Physics, M. V. Lomonosov Moscow State University, Moscow 119991, Russia.*
[2] *A. M. Obukhov Institute of Atmospheric Physics, Russian Academy of Sciences, Pyzhyovskiy per., Moscow 119017, Russia.*
E-mail: [a] *achulichkov@gmail.com,* [b] *sandratsy@list.ru.*

The classification of atmospheric signals was based on natural infrasound signals that were operated at Fairbanks, Alaska and Windless Bight, Antarctica from 1980 to 1983. The data files contained five subdirectories titled: «AIW» for auroral infrasonic waves, «MAW» for mountain associated waves, «VOL» for volcanic infrasound, «Microbarom» for microbaroms and «BombTest» for the 1980 Chinese nuclear test. The theory of testing statistical hypothesis was used for classification. The possibility of class separate was analyzed. It is shown that signals from used data typical for volcanic infrasound and the nuclear test are properly separate from typical signals for auroral infrasonic waves, mountain associated waves and microbaroms.

Key words: data analysis, mathematical modeling, signal shape, testing.
PACS: 02.50.Le.
Received 9 December 2011.

English version: *Moscow University Physics Bulletin* 2(2012).

Сведения об авторах

1. Чуличков Алексей Иванович — докт. физ.-мат. наук, профессор; тел.: (495) 939-41-78; e-mail: achulichkov@gmail.com.
2. Цыбульская Надежда Дмитриевна — аспирант; тел.: (495) 939-41-78; e-mail: sandratsy@list.ru.
3. Куличков Сергей Николаевич — зам. директора ИФА РАН.