УДК 681.513

А. В. Лапко, В. А. Лапко, А. В. Саренков

ГИБРИДНЫЙ АЛГОРИТМ РАСПОЗНАВАНИЯ ОБРАЗОВ И ЕГО СВОЙСТВА

Рассматривается методика синтеза и анализа гибридных алгоритмов распознавания образов, обеспечивающие эффективное использование априорных сведений о виде решающих функций и информации обучающих выборок. Исследуются их свойства аналитически и методом статистического моделирования.

Ключевые слова: распознавание образов, непараметрическая статистика, гибридные алгоритмы, асимптотические свойства.

При решении задач распознавания образов различают два типа исходной информации: априорные сведения о виде уравнения разделяющей поверхности и обучающая выборка, составленная из значений признаков классифицируемых объектов и соответствующих им «указаний учителя». Известные подходы к синтезу решающего правила классификации ориентированы в основном на определенный тип исходных данных, что при условиях, отличающихся от априорных предположений, приводит к снижению их эффективности. Так, в параметрических алгоритмах за основу принимаются сведения о виде уравнения разделяющей поверхности (ее аналитический вид), то для непараметрических процедур достаточно знание лишь ее качественных характеристик и информации обучающей выборки.

Для решения проблемы эффективного использования априорной информации предлагаются гибридные системы распознавания образов, которые обеспечивают сочетание в обобщенном решающем правиле классификации преимуществ параметрических и локальных методов аппроксимаций, основанных на оценках плотности вероятности типа Розенблатта–Парзена [1].

Синтез гибридного алгоритма распознавания образов. Пусть исходную информацию при решении двухальтернативной задачи распознавания образов составляют обучающая выборка $V = \left(x^i, \sigma(x^i),\, i = \overline{1,n}\right)$ и априорные сведения $F_{12}(x,\alpha)$ о виде уравнения разделяющей поверхности $f_{12}(x)$ между классами Ω_1, Ω_2 в пространстве $x \in R^k$. Знание $F_{12}(x,\alpha)$ предполагает наличие решающего правила классификации

$$m_{12}^F : \begin{cases} x \in \Omega_1, & \text{если } \mathrm{F}_{12}(x,\alpha) \le 0, \\ x \in \Omega_2, & \text{если } \mathrm{F}_{12}(x,\alpha) > 0, \end{cases} \tag{1}$$

по тем или иным причинам не удовлетворяющего исследователя. Информация обучающей выборки V формируется на основании данных о значениях признаков x классифицируемых объектов и соответствующих им «указаний учителя»

$$\sigma(x) = \begin{cases} -1, \text{если } x \in \Omega_1, \\ \ \ 1, \text{если } x \in \Omega_2. \end{cases} \tag{2}$$

Для использования в полном объеме априорной информации $\left(F_{12}(x,\alpha), V\right)$ воспользуемся принципами гибридного моделирования, которые обеспечивают сочетание в обобщенном решающем правиле классификации преимущества параметрических и локальных методов аппроксимации.

Для этого определим параметры α уравнения разделяющей поверхности $F_{12}(x,\alpha)$ решающего правила (1) из условия минимума эмпирической ошибки распознавания образов

$$\bar{\rho}(\alpha) = n^{-1} \sum_{j=1}^{n} 1\left(\sigma(x^j), \bar{\sigma}(x^j)\right), \tag{3}$$

где индикаторная функция

$$1\left(\sigma(x^j), \bar{\sigma}(x^j)\right) = \begin{cases} 1, & \text{если } \sigma(x^j) \ne \bar{\sigma}(x^j), \\ 0, & \text{если } \sigma(x^j) = \bar{\sigma}(x^j); \end{cases}$$

$\bar{\sigma}(x^j)$ – «решение» правила (1) о принадлежности ситуации x^j к тому или иному классу.

По результатам вычислительного эксперимента сформируем выборку расхождений $V_1 = \left(x^i, q(x^i),\, i = \overline{1,n}\right)$ между «решениями» $\bar{\sigma}(x^i)$ правила (1) и «указаниями учителя» $\sigma(x^i)$ из обучающей выборки *V*. При этом значения функции расхождений

$$q(x^i) = \begin{cases} 0 \ \forall\ \sigma(x^i) = \bar{\sigma}(x^i), \\ \left(\mathrm{F}_{12}(x^i, \bar{\alpha}) + \Delta\right) \ \forall\ \bar{\sigma}(x^i) = -1 \ и\ \sigma(x^i) = 1, \\ -(\mathrm{F}_{12}(x^i, \bar{\alpha}) + \Delta) \ \forall\ \bar{\sigma}(x^i) = 1 \ и\ \sigma(x^i) = -1. \end{cases}$$

При наличии ошибки, функция расхождения принимает значение обратное по знаку уравнения разделяющей поверхности $F_{12}(x,\alpha)$ и превышает его на величину Δ. Например, если ситуация x^i принадлежит второму классу ($\sigma(x^i) = 1$), а в соответствии с решающим правилом (1) классифицируемый объект с признаками $x^i \in \Omega_1$, т. е. $F_{12}(x^i, \alpha) < 0$, то значение функции расхождения в ситуации x^i соответствует $q(x^i) = F_{12}(x^i, \bar{\alpha}) + \Delta$.

Восстановление функции $q(x)$ по выборке V_1 осуществляется на основе непараметрической регрессии [2]:

$$\bar{q}(x) = \frac{\sum_{i=1}^{n} q(x^i) \beta_i(x)}{\sum_{i=1}^{n} \beta_i(x)},$$

$$\beta_i(x) = \prod_{\nu=1}^{k} \Phi\left(\frac{x_\nu - x_\nu^{\ i}}{c_\nu}\right), \tag{4}$$

где $\Phi(\cdot)$ – ядерная функция, удовлетворяющая условиям положительности, симметричности и нормированности [1].

Тогда гибридный алгоритм классификации запишется в виде

$$\bar{m}_{12}(x) : \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x) \le 0, \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x) > 0, \end{cases} \tag{5}$$

где

$$\bar{f}_{12}(x) = F_{12}(x, \bar{\alpha}) + \bar{q}(x). \tag{6}$$

Оптимизация алгоритма (5) по параметрам размытости c_v, $v=\overline{1,k}$ ядерных функций $\Phi(\cdot)$ и Δ осуществляется из условия минимума статистической оценки ошибки распознавания образов типа (3).

Меняя вид функции $q(x)$, обеспечивающей коррекцию $F_{12}(x,\overline{\alpha})$, можно получить семейство гибридных решающих правил [3].

Асимптотические свойства гибридных решающих функций. Рассмотрим асимптотические свойства гибридных уравнений разделяющих поверхностей $\overline{f}_{12}(x)$ $\forall\ x \in R^1$ при известной совместной плотности вероятности $p(x)$ распределения x в классах Ω_1, Ω_2.

Предположим, что $f_{12}(x)$, $F(x,\alpha)$, $p(x)$ ограничены и непрерывны со всем своими производными до второго порядка включительно. Эти условия, налагаемые на $f_{12}(x)$, $F(x,\alpha)$, $p(x)$, обозначим через G_2.

Теорема. Пусть: 1) уравнение разделяющей поверхности $f_{12}(x)$, ее аппроксимация $F(x,\alpha)$ и совместная плотность вероятности $p(x)$ распределения x в классах удовлетворяют условиям G_2; 2) ядерные функции $\Phi(\mathrm{u}) \geq 0$ в непараметрической статистике (4) являются симметричными и нормированными, причем значение $\int u^m \Phi(u)\, du$ при $m>2$ ограниченно и равно 1 при $m=2$; 3) последовательности $c=c(n)\to 0$, $\Delta=\Delta(n)\to 0$ при $n\to\infty$, а $nc\to\infty$.

Тогда гибридная модель $\overline{f}_{12}(x)$ (6) обладает свойствами асимптотической несмещенности

$$M\left(f_{12}(x)-\overline{f}_{12}(x)\right) \sim \frac{\left(q(x)p(x)\right)^{(2)}}{2p(x)}c^2+\Delta$$

и сходимости в среднеквадратическом

$$M(f_{12}(x)-\overline{f}_{12}(x))^2 \sim \left(ncp(x)\right)^{-1} q^2(x)\left\|\Phi(\mathrm{u})\right\|^2 +$$
$$+c^4\left(\left(q(x)p(x)\right)^{(2)}\right)^2\left(4p^2(u)\right)^{-1}+$$
$$+\frac{\left(q(x)p(x)\right)^{(2)}}{p(x)}c^2\Delta+\Delta^2.$$

Здесь $\left(q(x)p(x)\right)^{(2)}$ – вторая производная по x произведения функций в скобках; M – знак математического ожидания; $\left\|\Phi(u)\right\|^2=\int\Phi^2(u)\,du$.

Справедливость приведенных утверждений определяет состоятельность статистики $\overline{f}_{12}(x)$.

Доказательство основано на методике, предложенной в работе [4] при исследовании гибридных моделей стохастических зависимостей.

Анализ результатов вычислительных экспериментов. Исследовалась двуальтернативная задача распознавания образов k-мерном пространстве признаков ($k=\overline{2,10}$). Законы распределения признаков в области первого класса формировались в соответствии с датчиком случайных чисел

$$x_v = m+\sigma\left(\sum_{i=1}^{p}\varepsilon^i-0{,}5p\right)\frac{6}{\sqrt{3p}},\ v=\overline{1,k},$$

при $p=12$; $m=3$; $\varepsilon\in[0;1]$ – случайная величина с равномерным законом распределения; $\sigma=1$.

Значения признаков второго класса генерировались с использованием датчиков случайных чисел

$$x_v = a+\varepsilon\left(b-a\right),$$

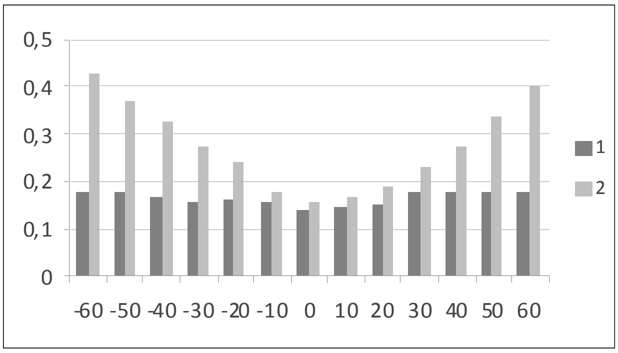
$$x_{v+1} = \left(x_v\right)^2-6x_v+10+\sigma\left(\sum_{i=1}^{p}\varepsilon^i-0{,}5p\right)\frac{6}{\sqrt{3p}},$$

где $v\in I_{н}$ – множество нечетных чисел меньших k, $a=1{,}5$, $b=4{,}5$.

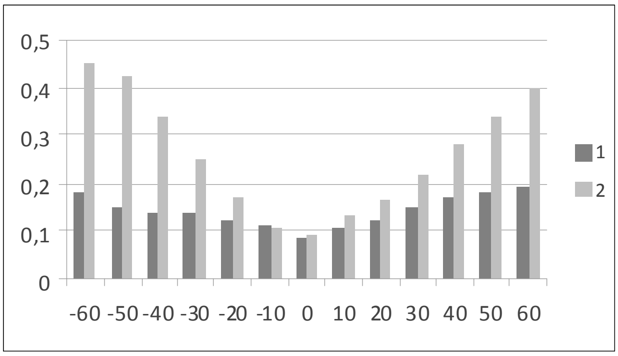
Исходные сведения о виде решающей функции представлялись полиномом

$$F(x,\alpha)=\sum_{v\in I_н}\left[x_{v+1}-\left(\left(x_v\right)^2-6x_v+10{,}8\right)\right].$$

Результаты вычислительных экспериментов представлены на рисунке. По горизонтальной оси отложено относительные (%) отклонения параметров решающей функции $F(x,\alpha)$ от оптимальных, по вертикальной оси – оценки вероятностей ошибок распознавания образов.



а



б

Зависимость оценок вероятностей ошибок распознавания образов гибридного (столбец 1) и параметрического (столбец 2) алгоритмов отклонений параметров исходной решающей функции от оптимальных. Размерность пространства признаков: $k=2$ (*а*); $k=4$ (*б*)

Из анализа данных на рисунке следует, что гибридный классификатор обладает значительной устойчивостью к отклонениям параметров исходной решающей функции от оптимальных. Например, для $k=2$ при отклонении параметров решающей функции от оптимальных на $\pm$ 60% оценка вероятности ошибки распознавания образов параметрического алгоритма больше гибридного на 63 %.

Таким образом, гибридные алгоритмы распознавания образов позволяют использовать информацию обучающих выборок и частичные априорные сведения о виде решающих функций на основе сочетания преимуществ параметрических и локальных аппроксимаций, основанных на ядерных оценках плотности вероятности типа Розенблатта–Парзена. Структура гибридных уравнений разделяющих поверхностей между классами в подобных системах формируется на основе параметрической ее аппроксимации, восстанавливаемой с учетом априорных сведений, и корректирующей ее функции непараметрического типа. Вид корректирующей функции и особенности исходной информации порождают семейство изучаемого класса систем.

Гибридные решающие функции обладают свойствами асимптотической несмещенности и состоятельности, устойчивы к отклонениям параметров исходного уравнения разделяющей поверхности от оптимальных их значений.

Библиографический список

1. Parzen, E. On estimation of a probability densiny function an mode / E. Parzen // Annals of Mathematical Statistics. 1962. Vol. 33. P. 1065–1076.

2. Надарая, Э. А. Непараметрические оценки кривой регрессии / З. А. Надарая // Тр. ВЦ АН ГССР. Вып. 5. 1965. С. 56–68.

3. Lapko, A. V. Hybrid Systems of Pattern Recognition / A. V Lapko, V. A. Lapko // Pattern recognition and image analysis. 2008. Vol. 18. № 1. P. 7–13.

4. Лапко, А. В. Имитационные модели неопределенных систем / А. В. Лапко. Новосибирск : Наука. 1993.

A. V. Lapko, V. A. Lapko, A. V. Sarenkov

HYBRID ALGORITHM OF PATTERN RECOGNITION AND ITS PROPERTIES

The technique of synthesis and analysis of hybrid algorithms pattern recognition, providing an effective utilization of aprioristic data on a kind of decision functions and information of training samples is considered. Their properties are investigated analytically by a method of statistical modeling.

Keywords: pattern recognition, nonparametric statistics, hybrid algorithms, asymptotic properties.

УДК 630*561.24:582.477.6

В. В. Шишов

МЕТОД ЭМПИРИЧЕСКИХ КАТЕГОРИАЛЬНЫХ КОРРЕЛОГРАММ И ЕГО ПРИМЕНЕНИЕ[1]

Предложен новый спектральный метод к анализу категориальных данных – метод эмпирических категориальных коррелограмм. На базе вычислительного эксперимента проводятся исследования на его статистическую устойчивость к различного рода шумовым воздействиям. Приводится пример анализа дендрохронологических данных на его основе.

Ключевые слова: спектральный анализ, категориальные данные.

В самом общем случае любая случайная функция натурального аргумента t (или временной ряд) может быть представлена в следующем виде [1]:

$$X(t) = A(t)\sin(\omega(t)t + \psi(t)).$$

В связи с этим, самая большая сложность, встречаемая в спектральном анализе, заключена в оценке параметров $A(t)$ – амплитуды колебаний, $\omega(t)$ – их частоты и $\psi(t)$ – фазового сдвига колебаний, которые являются также функциями времени. Проиллюстрируем эту проблему на примере оценки частоты колебаний.

Проанализируем взаимосвязи спектров циклических компонент, которые обладают примерно одинаковыми частотами. Отметим, что при определенных условиях циклические компоненты с близкими частотами будут линейно-независимыми, в частности, линейные корреляции между ними будут равны 0 [2]. Это вытекает из свойства, которое широко используется при преобразованиях Фурье, Хартли и различных модификаций этих методов [2]. А именно набор гармоник $\left\{\sin(2\pi \cdot t \cdot 1/j),\ \text{где}\ j = 2\pi/k,\ k - \text{целое}\right\}$ образует базис в бесконечно-мерном функциональном пространстве. На практике, в силу жестких ограничений (например, стационарность исходных временных рядов), накладываемых на использование чистого преобразования Фурье [2], широкое распространение получили методы (SSA, MTM или CWT), для которых такие ограничения не столь критичны.

Существует ряд классических примеров, которые показывают сложность определения при спектральном анализе истинных частот для временных рядов, встречаемых в различных областях естественных наук. Эти ряды, как правило, отличаются наличием нестационарных амплитуды, фазовых сдвигов и «колорированного» шума, кото-

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 09-05-00900-а, проект № 09-04-00803-а).