РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2006, том 42, № 6

УДК 681.513

СИНТЕЗ НЕЛИНЕЙНЫХ НЕПАРАМЕТРИЧЕСКИХ РЕШАЮЩИХ ПРАВИЛ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ОБРАЗОВ *

В. А. Лапко [1], А. Н. Капустин [2]

[1] *Институт вычислительного моделирования СО РАН, г. Красноярск*
E-mail: lapko@icm.krasn.ru
[2] *Красноярский государственный университет, г. Красноярск*

Предлагается методика построения нелинейных непараметрических алгоритмов распознавания образов, обеспечивающих эффективное решение задач классификации в условиях малых выборок. Проводится анализ их свойств по результатам вычислительных экспериментов.

Введение. Использование математического аппарата теории классификации и методов непараметрической статистики – одно из перспективных научных направлений исследования систем при априорной неопределенности. Его практическая значимость состоит в возможности создания типовых информационных средств, адаптируемых к условиям функционирования систем различной природы [1, 2].

Однако по мере усложнения изучаемых объектов появляются методологические и вычислительные трудности применения традиционных непараметрических алгоритмов, что особо наблюдается при обработке массивов разнотипных данных большой размерности. Подобные условия часто встречаются при исследовании медико-биологических, экологических и социально-экономических процессов, когда количество параметров, характеризующих их состояние, и объем исходных экспериментальных данных соизмеримы [3].

Перспективное направление обхода возникающих проблем малых выборок при моделировании неопределенных систем основывается на последовательных процедурах формирования решений, которые предполагают разбиение исходной задачи на ряд взаимосвязанных более простых задач. Такая схема используется, например, в методе комитетов [4], методе группового учета аргументов (МГУА) [5] и синтезе многоуровневых систем обработки информации [6].

В данной работе с позиции последовательных процедур принятия решений и принципов коллективного оценивания предлагаются статистические

* Работа выполнена при поддержке Совета по грантам Президента РФ (гранты № МД-2130.2005.9, № НШ-3428.2006.9).

модели распознавания образов, представляющие собой семейство частных решающих функций, организация которых в нелинейном классификаторе осуществляется с помощью методов непараметрической статистики. Частные решающие функции формируются на основе однородных частей обучающей выборки, которые удовлетворяют одному или нескольким требованиям: наличию однотипных признаков, пропуску данных, возможности декомпозиции исходных признаков на группы в соответствии со спецификой задачи распознавания образов. Это порождает широкий круг постановок задач синтеза непараметрических алгоритмов классификации. При интеграции частных решающих функций используются непараметрические оценки оптимальных байесовских алгоритмов распознавания образов.

Синтез нелинейных непараметрических решающих правил. Рассмотрим методику построения нелинейного непараметрического классификатора на примере двухальтернативной задачи распознавания образов в пространстве непрерывных признаков.

Пусть $V = (x^i, \sigma(x^i),\ i = \overline{1,n})$ – обучающая выборка объема n, составленная из значений признаков $x^i = (x_1^i, x_2^i, \ldots, x_k^i)$ классифицируемых объектов и соответствующих «указаний учителя» об их принадлежности к одному из двух классов:

$$\sigma(x^i) = \begin{cases} -1, & \text{если } x^i \in \Omega_1; \\ 1, & \text{если } x^i \in \Omega_2. \end{cases}$$

Причем отношение размерность/объем выборки соизмеримо с единицей.

Условные плотности вероятности распределения значений признаков x в области определения классов неизвестны.

Идея предлагаемого подхода к решению задачи распознавания образов в данных условиях состоит в выполнении следующих действий.

1. В соответствии с особенностями задачи классификации формируются наборы признаков $(x(t),\ t = \overline{1,T})$, и на этой основе осуществляются декомпозиции исходной выборки $V = (x^i, \sigma(x^i),\ i = \overline{1,n})$ на однородные части $V(t) = (x^i(t),\ \sigma(x^i(t)),\ i = \overline{1,n}),\ t = \overline{1,T}$.

2. По полученным данным строятся решающие правила

$$m_t(x(t)):\ \begin{cases} x \in \Omega_1, & \text{если } \bar{f}_{12}(x(t)) \le 0; \\ x \in \Omega_2, & \text{если } \bar{f}_{12}(x(t)) > 0, \quad t = \overline{1,T}. \end{cases} \tag{1}$$

В качестве оценок частных решающих функций $f_{12}(x(t))$ между классами в пространстве признаков $x_\nu,\ \nu \in I_t \subset I = (\nu = \overline{1,k})$, используются непараметрические статистики

$$\bar{f}_{12}(x(t)) = \left[n \prod_{\nu \in I_t} c_\nu \right]^{-1} \sum_{i=1}^{n} \sigma(x^i(t)) \prod_{\nu \in I_t} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu} \right), \quad t = \overline{1,T}, \tag{2}$$

где $\Phi(\cdot)$ – ядерные функции, удовлетворяющие условиям положительности, симметричности, нормированности и имеющие конечные центральные моменты [1, 7]. Оптимизация частных решающих правил (1) по коэффициентам

размытости ядерных функций c_ν, $\nu \in I_t$, осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов

$$\overline{\rho}(t) = \frac{1}{n}\sum_{j=1}^{n} 1(\sigma(x^j(t)), \overline{\sigma}(x^j(t))), \quad t = \overline{1,T},$$

$$1(\sigma(j), \overline{\sigma}(j)) = \begin{cases} 0, & \text{если } \sigma(x^j(t)) = \overline{\sigma}(x^j(t)); \\ 1, & \text{если } \sigma(x^j(t)) \neq \overline{\sigma}(x^j(t)), \end{cases}$$

где $\overline{\sigma}(x^j(t))$ – решение о принадлежности ситуации $x^j(t)$ к одному из двух классов с помощью алгоритма (1). При формировании решения $\overline{\sigma}(x^j(t))$ ситуация $x^j(t)$ исключается из процесса обучения в непараметрической статистике (2).

3. С использованием непараметрических оценок решающих функций (2) формируется обучающая выборка

$$(\bar{f}_{12}(x^i(1)), \bar{f}_{12}(x^i(2)), \ldots, \bar{f}_{12}(x^i(T)), \sigma(x^i), \quad i = \overline{1,n})$$

и строится решающее правило в пространстве значений $\bar{f}_{12}(x(t))$, $t = \overline{1,T}$:

$$m(f_{12}(x(t))): \begin{cases} x \in \Omega_1, & \text{если } F_{12}(f_{12}(x(t))) \le 0; \\ x \in \Omega_2, & \text{если } F_{12}(f_{12}(x(t))) > 0, \end{cases} \tag{3}$$

где непараметрическая оценка обобщенной решающей функции между классами имеет вид

$$\overline{F}_{12}(f_{12}(x(t))) = \left[n \prod_{\nu=1}^{T} c_\nu \right]^{-1} \sum_{i=1}^{n} \sigma(x^i) \prod_{\nu=1}^{T} \Phi\left(\frac{f_{12}(x(t)) - \bar{f}_{12}(x^i(t))}{c_\nu} \right). \tag{4}$$

Структура нелинейного непараметрического алгоритма распознавания образов представлена на рис. 1. На первом уровне структуры системы клас-

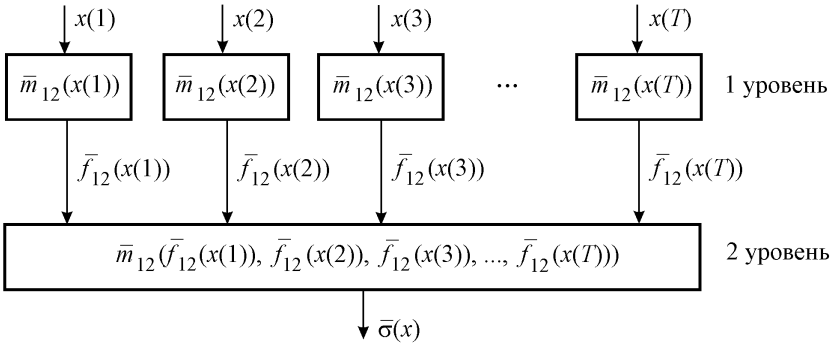

Рис. 1. Двухуровневая структура нелинейного непараметрического классификатора

сифицируемая ситуация x преобразуется в значения непараметрических оценок $\bar{f}_{12}(x(t))$, $t=\overline{1,T}$, в пространстве которых принимается решение $\overline{\sigma}(x)$ с помощью правила (3) о принадлежности ситуации x к тому или иному классу.

Предлагаемый алгоритм классификации обеспечивает не только эффективное решение задач распознавания образов в условиях малых выборок, но и позволяет учитывать априорные сведения о виде частных решающих функций.

Ближайшими аналогами данного алгоритма являются метод комитетов [4] и алгебраический подход [8].

Метод комитетов основан на линейном преобразовании семейства упрощенных решающих правил. Его недостаток заключается в потере информации при переходе от исходного вектора описания классифицируемых объектов к его булевому представлению в процессе формирования обобщенного уравнения разделяющей поверхности. Несмотря на близость их структуры, отличие рассматриваемого подхода состоит в нелинейном преобразовании частных решающих функций в пространстве непрерывных значений этих функций с помощью методов непараметрической статистики.

Общим для алгебраического и предлагаемого подходов является использование нелинейных процедур последовательного конструирования новых эффективных алгоритмов распознавания образов. В алгебраическом подходе алгоритм распознавания образов представляется в виде двух операторов: распознающего и решающего правил. Распознающий оператор определяет меру близости контрольных объектов с каждым классом. С распознающими операторами исходных алгоритмов распознавания образов можно производить алгебраические операции (сложение, умножение и умножение на число), что обеспечивает расширение их семейства, содержащего корректный алгоритм решения задачи распознавания образов. Пороги, константы подобных алгоритмов и порядок алгебраического замыкания определяются на основе специальной обработки априорной информации.

Отличие нелинейных непараметрических решающих правил состоит в особенностях преобразования мер близости между контрольной ситуацией и классами, а также в используемых математических средствах формирования обобщенного решения. Применение предлагаемых непараметрических алгоритмов позволяет обойти проблемы выбора вида преобразований исходной и промежуточной информации при формировании обобщенного решающего правила. При этом значительно сокращается размерность задачи оптимизации непараметрических алгоритмов, которая сводится к выбору коэффициентов размытости ядерных функций в статистиках (2), (4) из условия минимума эмпирических ошибок распознавания образов.

Исследование свойств нелинейного непараметрического алгоритма распознавания образов. На основании данных вычислительного эксперимента сравнивается эффективность нелинейного непараметрического классификатора решающих правил с хорошо зарекомендовавшим себя на практике традиционным непараметрическим алгоритмом распознавания образов [1, 2].

Традиционный непараметрический алгоритм основывается на решающей функции типа (2), определенной в пространстве признаков $x=(x_1, x_2, \ldots, x_k)$.

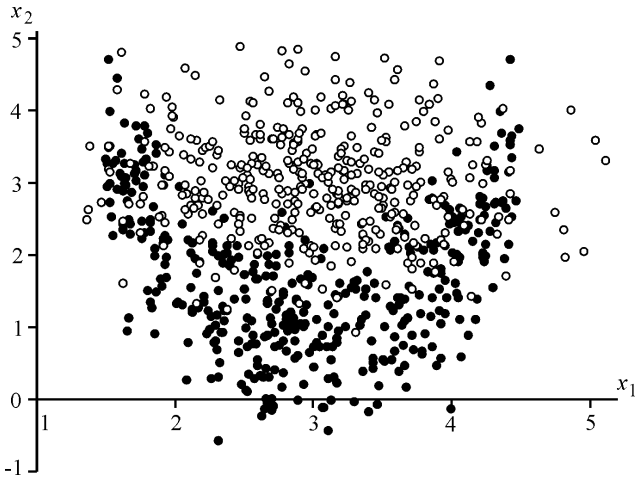

Рис. 2. Элементы исходной выборки в пространстве двух признаков (x_1, x_2) при $n = 600$ (• – ситуация первого класса, ∘ – ситуация второго класса)

Исследования осуществлялись при решении двухальтернативной задачи распознавания образов в k-мерном пространстве признаков. Законы распределения признаков в области первого класса формировались в соответствии с датчиками случайных чисел

$$x_{\nu} = a + \varepsilon (b - a),$$

$$x_{\nu+1} = (x_{\nu})^2 - 6x_{\nu} + 10 + \sigma_1 \left(\sum_{i=1}^{p_1} \varepsilon^i - 0{,}5 p_1 \right) \frac{6}{\sqrt{3p_1}}, \quad \nu \in I_{\text{н}},$$

где $a = 1{,}5$, $b = 4{,}5$, $p_1 = 5$ – параметры распределений; $\sigma_1 = 0{,}7$ – среднеквадратичное отклонение; $\varepsilon \in [0;1]$ – случайная величина с равномерным законом распределения; $I_{\text{н}} = (1, 3, 5, \ldots)$ – множество нечетных чисел, меньших k.

Признаки второго класса генерировались с нормальным законом

$$x_{\nu} = m + \sigma_2 \left(\sum_{i=1}^{p_2} \varepsilon^i - 0{,}5\, p_2 \right) \frac{6}{\sqrt{3p_2}}, \quad \nu = \overline{1, k},$$

при $p_2 = 5$, $\sigma_2 = 0{,}7$, $m = 3$.

Априорные вероятности классов $P_1 = P_2 = 0{,}5$. Элементы исходной выборки в пространстве двух признаков (x_1, x_2) представлены на рис. 2, а их преобразования в пространство значений частных решающих функций $(\bar{f}_{12}(x^i(1)), \bar{f}_{12}(x^i(2)), \sigma(x^i), \ i = \overline{1, n})$ при $T = 2$ и $k = 4$ иллюстрирует рис. 3.

Основные характеристики методики статистического моделирования:

– формирование многомерной случайной величины осуществляется на основе независимой генерации ее координат;

– вычислительный эксперимент при фиксированных условиях исследований производился N раз по контрольной выборке, полученной с помощью приведенных выше датчиков случайных чисел при $\sigma_1 = \sigma_2 = 0{,}6$, $p_1 = p_2 = 5$, $n = 2000$, $P_1 = P_2 = 0{,}5$.

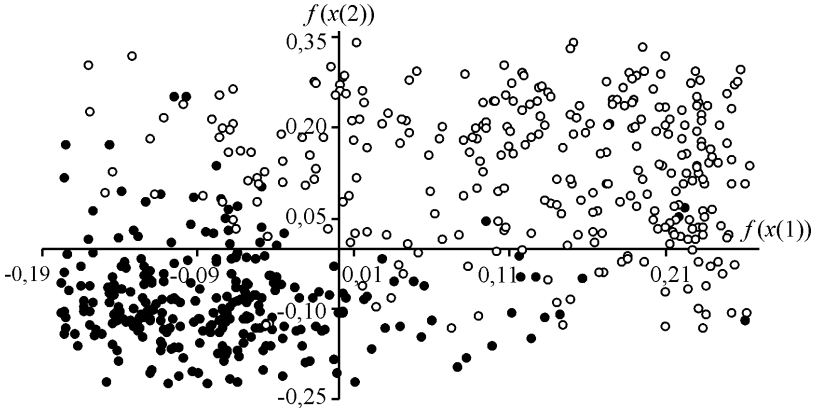

Рис. 3. Элементы выборки, используемые при синтезе нелинейного непараметрического алгоритма распознавания образов при $x(1) = (x_1, x_2)$, $x(2) = (x_3, x_4)$, $n = 600$ (• – ситуация первого класса, ◦ – ситуация второго класса)

– достоверность различия эмпирических оценок вероятностей ошибок распознавания образов сравниваемых методов рассчитывалась в соответствии с критерием Смирнова.

Пусть D_N – максимальное расхождение эмпирических функций распределения оценок вероятностей ошибок распознавания образов, соответствующих нелинейному непараметрическому решающему правилу и традиционному классификатору ядерного типа. Тогда показатели эффективности сравниваемых алгоритмов не отличаются, если выполняется соотношение

$$D_N < \sqrt{\frac{-\ln(\beta/2)}{N}},$$

где β – вероятность отвергнуть проверяемое утверждение ($\beta = 0{,}05$).

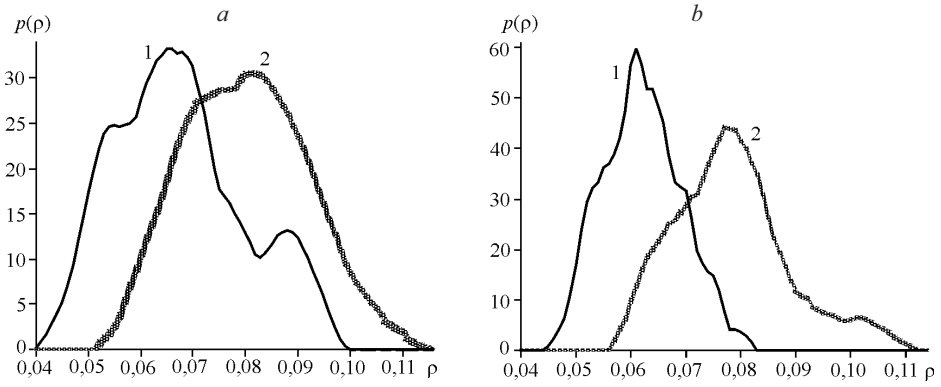

Рис. 4. Оценки плотностей вероятностей эмпирических ошибок распознавания образов нелинейного непараметрического решающего правила при $T = 2$ (кривые 1) и традиционного непараметрического классификатора (кривые 2). Условия эксперимента: $k = 4$, $N = 50$; объем обучающей выборки $n = 100$ (*a*), $n = 200$ (*b*)

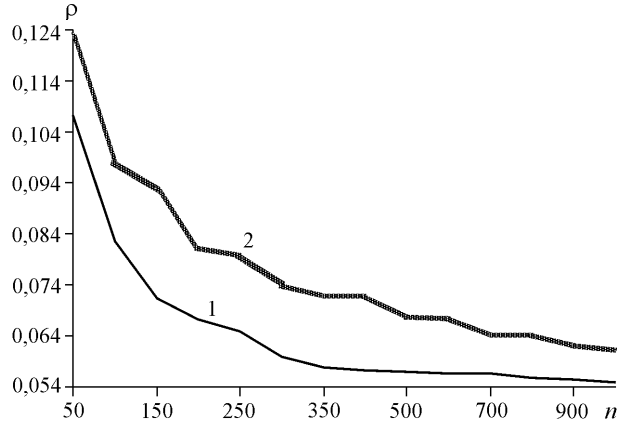

Рис. 5. Зависимости оценок вероятностей ошибок распознавания образов от объема обучающей выборки при $k = 4$ и $N = 50$. Кривая 1 соответствует нелинейному непараметрическому решающему правилу при $T = 2$, кривая 2 – традиционному непараметрическому классификатору

Анализ результатов вычислительного эксперимента показал, что законы распределения оценок вероятностей ошибок классификации сравниваемых алгоритмов распознавания образов являются симметричными (рис. 4). Поэтому средние значения оценок исследуемых показателей эффективности близки к их наиболее вероятным значениям.

На рис. 5 приведены зависимости средних значений оценок вероятностей ошибок распознавания образов нелинейного непараметрического классификатора при $T = 2$ и традиционного непараметрического алгоритма при $k = 4$ от объема обучающей выборки n.

Установлено достоверное преимущество предлагаемого алгоритма перед традиционным. Данная закономерность сохраняется для различных объемов обучающих выборок при значении $\beta = 0{,}05$.

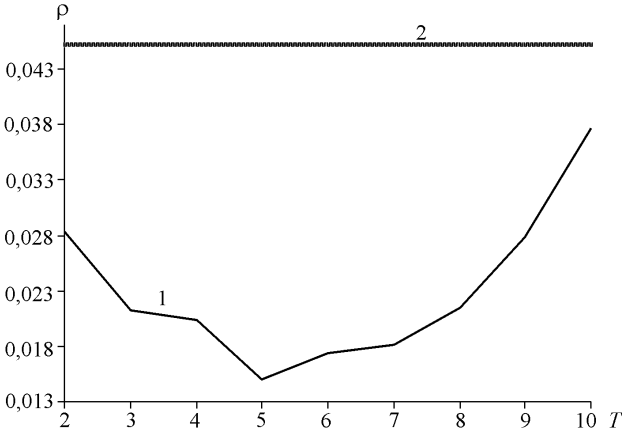

Рис. 6. Зависимости оценок вероятностей ошибок распознавания образов от количества частных решающих правил T при $n = 100$, $N = 50$, $k = 10$. Кривая 1 соответствует нелинейному непараметрическому решающему правилу, кривая 2 – традиционному непараметричскому классификатору

Обнаружен экстремальный характер зависимости показателей эффективности нелинейного непараметрического классификатора от количества T частных решающих правил (рис. 6). Вследствие роста размерности k признаков классифицируемых объектов его преимущество при оптимальных значениях T перед традиционными непараметрическими алгоритмами возрастает. Отношение средних значений оценок алгоритмов вероятности ошибки достигает трех на контрольных выборках, что особенно проявляется при малых объемах экспериментальных данных. Причем это отношение на обучающих выборках возрастает до 10.

Заключение. Нелинейные непараметрические алгоритмы распознавания образов являются эффективным средством решения задач классификации в условиях малых обучающих выборок. Их применение обеспечивает значительное снижение ошибки распознавания образов на контрольных выборках в 1,5–3 раза по сравнению с традиционным непараметрическим классификатором.

Перспективы развития предлагаемого подхода связаны с его применением в задачах классификации в условиях разнотипной информации и неоднородных выборок, получаемых в результате заполнения пропусков данных.

СПИСОК ЛИТЕРАТУРЫ

1. **Медведев А. В.** Непараметрические системы адаптации. Новосибирск: Наука, 1983.
2. **Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.** Непараметрические системы классификации. Новосибирск: Наука, 2000.
3. **Лапко А. В., Поликарпов Л. С., Манчук В. Т. и др.** Автоматизиция научных исследований в медицине (по данным популяционных обследований). Новосибирск: Наука, 1996.
4. **Мазуров Вл. Д.** Метод комитетов в задачах оптимизации и классификации. М.: Наука, 1990.
5. **Ивахненко А. Г., Чаинская В. А., Ивахненко Н. А.** Непараметрический комбинаторный алгоритм МГУА на операторах поиска аналогов // Автоматика. 1990. № 5. С. 14.
6. **Лапко А. В., Ченцов С. В., Крохов С. И., Фельдман Л. А.** Обучающиеся системы обработки информации и принятия решений. Новосибирск: Наука, 1996.
7. **Хардле В.** Прикладная непараметрическая регрессия. М.: Мир, 1993.
8. **Журавлев Ю. И.** Об алгебраическом подходе к решению задач распознавания образов или классификации // Проблемы кибернетики. 1978. № 33. С. 5.

Поступила в редакцию 5 декабря 2005 г.