УДК 681.513

А. В. Лапко, В. А. Лапко, А. Н. Хлопов

НЕПАРАМЕТРИЧЕСКИЙ АЛГОРИТМ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ СТАТИСТИЧЕСКИХ ДАННЫХ

Предлагается непараметрический алгоритм автоматической классификации статистических данных, основу которого составляют оценки плотности вероятности парзеновского типа. Применение алгоритма позволяет выделять компактные множества точек, соответствующих одномодальным фрагментам плотности вероятности.

Ключевые слова: *непараметрическая статистика, автоматическая классификация, распознавание образов, плотность вероятности.*

Методы автоматической классификации статистических данных широко используются при разработке математического обеспечения аппаратно-программных комплексов технического зрения, а также при создании систем обработки данных дистанционного зондирования Земли. Среди алгоритмов автоматической классификации, основанных на явном определении класса, следует отметить те, которые ориентированы на обнаружение множества объектов, соответствующих одномодальным фрагментам совместной плотности вероятности в заданном пространстве признаков [1, 2]. Определение класса связано с понятием закономерности в вероятностном смысле и имеет прикладную направленность в задачах синтеза структуры сложных систем, аппроксимации неоднозначных стохастических зависимостей.

Широкое распространение получили методы, основанные на оценивании смеси плотностей вероятности при неизвестном количестве классов и последующем анализе с помощью оптимизационных алгоритмов [3]. Выделение абстрактных образов рассматривается как поиск локальных экстремумов — максимумов непараметрической оценки плотности вероятности смеси. Однако реализация этого метода требует решения большого количества оптимизационных задач, равного числу классифицируемых объектов.

В настоящей статье задача автоматической классификации статистических данных решается формализованно в рамках задачи распознавания образов с помощью итерационной процедуры последовательного восстановления непараметрической оценки уравнения разделяющей поверхности между классами, соответствующими одномодальным фрагментам совместной плотности вероятности.

Базовый алгоритм классификации. Пусть имеется выборка $V = x^i, i = \overline{1, n}$, составленная из значений признаков $x = x_v, v = \overline{1, k}$, классифицируемых объектов. Необходимо разбить выборку V на группы компактных точек (классов), соответствующих одномодальным фрагментам совместной плотности вероятности $p(x)$. Априори количество M классов и вид $p(x)$ неизвестны.

При синтезе данного базового алгоритма полагается, что минимальное расстояние между элементами класса Ω_j и области $\Omega_{\bar{j}} = \bigcup\limits_{\substack{\lambda=1,\\ \lambda\neq j}}^{M} \Omega_\lambda$ больше порогового значения d:

$$\min_{x^i, x^t} \max_{v=1,k} \left| x_v^i - x_v^t \right| > d, \quad x^i \in \Omega_j, \; x^t \in \Omega_{\bar{j}}. \tag{1}$$

Восстановим непараметрическое уравнение разделяющей поверхности между классом Ω_j и областью $\Omega_{\bar{j}}$, которое представляется следующим образом [4]:

$$\overline{f}_{j\bar{j}}(x) = \frac{1}{n\prod_{v=1}^{k} c_v} \sum_{i=1}^{n} \sigma(i) \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right), \tag{2}$$

где ядерные функции $\Phi(u)$ удовлетворяет условиям

$$0 \le \Phi(u) < \infty, \quad \int_{-\infty}^{+\infty} \Phi(u)\, du = 1, \quad \Phi(u) = \Phi(-u), \quad \int_{-\infty}^{+\infty} u^2\, \Phi(u)\, du = 1,$$

$$\int_{-\infty}^{+\infty} u^m\, \Phi(u)\, du < \infty \;\forall\; 0 \le m < \infty,$$

а $c_v = c_v(n)$, $v = \overline{1,k}$, — последовательности коэффициентов размытости ядерных функций, убывающие с увеличением n.

Для восстановления непараметрической оценки уравнения разделяющей поверхности (2) необходимо идентифицировать

$$\sigma(i) = \begin{cases} 1, & \text{если } x^i \in \Omega_j; \\ -1, & \text{если } x^i \in \Omega_{\bar{j}}, \end{cases}$$

и определить оптимальные значения $c_v = c_v(n)$, $v = \overline{1,k}$.

Определим коэффициенты размытости для $\overline{f}_{j\bar{j}}(x)$ исходя из условия минимума критерия

$$W_1(c) = \int_{-\infty}^{+\infty} \left(f_{j\bar{j}}(x) - \overline{f}_{j\bar{j}}(x)\right)^2 dx. \tag{3}$$

В выражении (3) статистика

$$\overline{f}_{j\bar{j}}(x) = \overline{P}_j\, \overline{p}_j(x) - \overline{P}_{\bar{j}}\, \overline{p}_{\bar{j}}(x)$$

является оценкой байесовского уравнения разделяющей поверхности [2]

$$f_{j\bar{j}}(x) = P_j\, p_j(x) - P_{\bar{j}}\, p_{\bar{j}}(x),$$

где P_j, $P_{\bar{j}}$ — априорные вероятности принадлежности значения x к j-му классу и области $\Omega_{\bar{j}}$, а $\overline{P}_j$, $\overline{P}_{\bar{j}}$ — их статистические оценки.

При синтезе $\overline{f}_{j\bar{j}}(x)$ используются непараметрические оценки плотности вероятности типа Розенблатта — Парзена [5]: например,

$$\overline{p}_j(x) = \frac{1}{n_j \prod_{v=1}^{k} c_v} \sum_{i \in I_j} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right),$$

где I_j — множество элементов j-го класса из выборки V, а n_j — их количество.

Преобразуем выражение (3):

$$W_1(c) = \int_{-\infty}^{+\infty} \left(P_j\, p_j(x) - P_{\bar{j}}\, p_{\bar{j}}(x) - \overline{P}_j\, \overline{p}_j(x) + \overline{P}_{\bar{j}}\, \overline{p}_{\bar{j}}(x) \right)^2 dx =$$

$$= \int_{-\infty}^{+\infty} \left[\left(P_j\, p_j(x) - \overline{P}_j\, \overline{p}_j(x) \right) - \left(P_{\bar{j}}\, p_{\bar{j}}(x) - \overline{P}_{\bar{j}}\, \overline{p}_{\bar{j}}(x) \right) \right]^2 dx =$$

$$= \int_{-\infty}^{+\infty} \left(P_j\, p_j(x) - \overline{P}_j\, \overline{p}_j(x) \right)^2 dx + \int_{-\infty}^{+\infty} \left(P_{\bar{j}}\, p_{\bar{j}}(x) - \overline{P}_{\bar{j}}\, \overline{p}_{\bar{j}}(x) \right)^2 dx -$$

$$-2 \int_{-\infty}^{+\infty} \left(P_j\, p_j(x) - \overline{P}_j\, \overline{p}_j(x) \right) \left(P_{\bar{j}}\, p_{\bar{j}}(x) - \overline{P}_{\bar{j}}\, \overline{p}_{\bar{j}}(x) \right) dx .$$

Так как в соответствии с постановкой задачи автоматической классификации классы не пересекаются, то третье слагаемое равно нулю.

Аналогичным образом определим среднеквадратический критерий

$$W_2(c) = \int_{-\infty}^{+\infty} \left(p(x) - \overline{p}(x) \right)^2 dx$$

расхождения между совместной плотностью вероятности

$$p(x) = P_j\, p_j(x) + P_{\bar{j}}\, p_{\bar{j}}(x)$$

и ее непараметрической оценкой

$$\overline{p}(x) = \overline{P}_j\, \overline{p}_j(x) + \overline{P}_{\bar{j}}\, \overline{p}_{\bar{j}}(x) .$$

Нетрудно показать, что $W_2(c)$ отличается от $W_1(c)$ только знаком третьего слагаемого, которое в соответствии с определением класса равно нулю.

Отсюда следует, что выбор оптимальных коэффициентов размытости ядерных функций в непараметрическом уравнении разделяющей поверхности (2) сводится к их определению согласно условию минимума статистической оценки критерия

$$W_2(c) = \left[\int_{-\infty}^{+\infty} \overline{p}^2(x)\, dx - 2 \int_{-\infty}^{+\infty} \overline{p}(x) p(x)\, dx + \int_{-\infty}^{+\infty} p^2(x)\, dx \right].$$

Так как третье его слагаемое не зависит от искомого параметра c, то, оценивая второе слагаемое в виде среднего значения $\overline{p}(x)$, получаем критерий

$$\overline{W}_2(c) = \frac{1}{n^2 \prod_{v=1}^{k} c_v^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{-\infty}^{+\infty} \left[\prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v} \right) \Phi\left(\frac{x_v - x_v^j}{c_v} \right) \right] dx -$$

$$-\frac{2}{n} \sum_{i=1}^{n} \left(\frac{1}{n \prod_{v=1}^{k} c_v} \sum_{\substack{j=1 \\ i \neq j}}^{n} \prod_{v=1}^{k} \Phi\left(\frac{x_v^i - x_v^j}{c_v} \right) \right),$$

минимизация которого позволяет найти оптимальные параметры статистики (2).

Таким образом, не зная вид $\overline{f}_{j\bar{j}}(x)$, можно определить ее параметры $\overline{c}_v = \overline{c}_v(n)$, $v = \overline{1, k}$.

Будем считать, что $\overline{c} < d$, где d — минимальное расстояние между классами Ω_1 и $\Omega_{\bar{1}}$. С учетом условия (1) для реализации базового алгоритма классификации необходимо выполнить следующие действия.

1. Выбрать из исходной выборки $V = x^i$, $i = \overline{1, n}$, точку x^i, в которой $p(x^i) \neq 0$, и отнести ее к первому классу, т.е. $x^i \in \Omega_1$ и $\sigma(i) = 1$.

2. Осуществить первый этап классификации точек, принадлежащих классу Ω_1, в соответствии с правилом

$$x^t \in \Omega_1 \text{ и } \sigma(t) = 1, \text{ если } \prod_{v=1}^{k} \frac{1}{\overline{c}_v} \Phi\left(\frac{x_v^t - x_v^i}{\overline{c}_v} \right) > 0,\ t \in I \setminus (i),\ I = i = \overline{1, n}. \qquad (4)$$

Справедливость правила (4) следует из условия $\overline{c}_v < d$, $v = \overline{1, k}$.

Обозначим множество номеров точек, принадлежащих в соответствии с правилом (4) к первому классу, через I_1^1, включая номер i.

3. Провести классификацию точек, принадлежащих классу Ω_1, по следующему правилу:

$$x^t \in \Omega_1 \text{ и } \sigma(t) = 1, \text{ если } \frac{1}{\left| I_1^1 \right|} \sum_{\gamma \in I_1^1} \sigma(\gamma) \frac{1}{\overline{c}_v} \prod_{v=1}^{k} \Phi\left(\frac{x_v^t - x_v^\gamma}{\overline{c}_v} \right) > 0,\ t \in I \setminus I_1^1,\ I = i = \overline{1, n},$$

где $\left| I_1^1 \right|$ — количество элементов множества I_1^1.

Обозначим через I_1^2 множество номеров точек, принадлежащих на втором и третьем шаге классификации к классу Ω_1.

4. Продолжить классификацию точек, принадлежащих классу Ω_1, по правилу

$$x^t \in \Omega_1 \text{ и } \sigma(t) = 1, \text{ если } \frac{1}{\left| I_1^2 \right|} \sum_{\gamma \in I_1^2} \sigma(\gamma) \frac{1}{\overline{c}_v} \prod_{v=1}^{k} \Phi\left(\frac{x_v^t - x_v^\gamma}{\overline{c}_v} \right) > 0,\ t \in I \setminus I_1^2.$$

5. Предложенную методику классификации продолжать до тех пор, пока на некотором $(S+1)$-м этапе в соответствии с правилом

$$x^t \in \Omega_1 \text{ и } \sigma(t) = 1, \text{ если } \frac{1}{\left| I_1^S \right|} \sum_{\gamma \in I_1^S} \sigma(\gamma) \frac{1}{\overline{c}_v} \prod_{v=1}^{k} \Phi\left(\frac{x_v^t - x_v^\gamma}{\overline{c}_v} \right) > 0$$

к первому классу не будет отнесена ни одна из точек x^t, $t \in I \setminus I_1^S$.

Таким образом, множество точек x^i, $i \in I_1^S$, образуют первый класс Ω_1, а x^i, $i \in I \setminus I_1^S$, — объединение остальных классов Ω_j, $j = \overline{2, M}$.

При этом непараметрическое уравнение разделяющей поверхности между классом Ω_1 и областью $\Omega_{\bar{1}}$ имеет вид

$$\bar{f}_{1\bar{1}}(x) = \frac{1}{n}\sum_{i=1}^{n}\sigma(i)\prod_{v=1}^{k}\frac{1}{\bar{c}_v}\Phi\left(\frac{x_v - x_v^i}{\bar{c}_v}\right),$$

где

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

Аналогичным образом можно выделить точки, принадлежащие второму классу и всем остальным, если $d < \bar{c}_v$, $v = \overline{1,k}$.

Ближайшим аналогом базового алгоритма классификации является алгоритм „Форель“ [6].

Обобщенный алгоритм классификации. Если расстояние между классами $d = 0$, для решения задачи автоматической классификации предлагается выполнить следующие действия.

1. Задать некоторое значение непараметрической оценки $p_1 > 0$ совместной плотности вероятности $\bar{p}(x)$ и из исходной выборки $V = x^i$, $i = \overline{1,n}$, выделить множество точек $V_1 = x^i : \bar{p}(x^i) > p_1$, $i = \overline{1,n}$, со значением $\bar{p}(x)$, превышающим p_1.

Множество V_1 может содержать точки, принадлежащие центру Ω_j^1 некоторого класса Ω_j и области $\Omega_{\bar{j}}^1 = \bigcup_{\substack{t=1 \\ t\neq j}}^{M}\Omega_t^1$, расстояние между которыми $d_1 > 0$.

2. Используя базовый алгоритм автоматической классификации, провести декомпозицию выборки V_1. Если d_1 больше хотя бы одного из значений коэффициентов $\bar{c}_v$, $v = \overline{1,k}$, непараметрической оценки плотности вероятности $\bar{p}(x)$, то в соответствии с методикой, принятой для базового алгоритма, будут обнаружены множества $V_1(j)$, $V_1(\bar{j})$ точек, определяющих центры Ω_j^1, $\Omega_{\bar{j}}^1$ класса Ω_j и области $\Omega_{\bar{j}}$. Для идентификации остальных точек j-го класса перейти к п. 3.

Если центры классов не обнаружены, то необходимо увеличить значение p_1 на величину Δp и перейти к п. 1. В этом случае расстояние d_1 между центрами j-го класса и области $\Omega_{\bar{j}}^1$ увеличится и вероятность того, что $d_1 > \bar{c}_v$, $v = \overline{1,k}$, повысится.

3. Сформировать обучающую выборку $x^i, \sigma(i), i \in I_1$, здесь I_1 — множество номеров точек из V_1, а

$$\sigma(i) = \begin{cases} 1, & \text{если } x^i \in V_1(j), \\ -1, & \text{если } x^i \in V_1(\bar{j}). \end{cases}$$

4. Построить непараметрическое решающее правило распознавания образов

$$\bar{m}_{j\bar{j}}^1(x): \begin{cases} x \in \Omega_j, \text{ если } \bar{f}_{j\bar{j}}^1(x) > 0; \\ x \in \Omega_{\bar{j}}, \text{ если } \bar{f}_{j\bar{j}}^1(x) \le 0, \end{cases}$$

где

$$\overline{f}^{1}_{j\overline{j}}(x)=\frac{1}{|I_1|}\sum_{i\in I_1}\sigma(i)\prod_{v=1}^{k}\frac{1}{\overline{c}_v}\Phi\left(\frac{x_v-x_v^i}{\overline{c}_v}\right).$$

5. В соответствии с правилом $\overline{m}^{1}_{j\overline{j}}(x)$ осуществить классификацию оставшихся точек x^i, $i\in I\setminus I_1$, из исходной статистической выборки. Нетрудно заметить, что j-му классу будут принадлежать новые точки x^i, находящиеся в $\overline{c}$-окрестности граничных точек множества $V_1(j)$.

6. По результатам классификации расширить обучающую выборку $x^i,\sigma(i),i\in I_1$, где $I_1=I_1\bigcup R$; $R=R_j\bigcup R_{\overline{j}}$ — множество номеров точек, принадлежащих на шаге 5 к классу Ω_j и области $\Omega_{\overline{j}}$. При этом

$$V_1(j)=V_1(j)\bigcup x^i,\ i\in R_j\ ;\quad V_1(\overline{j})=V_1(\overline{j})\bigcup\ x^i,\ i\in R_{\overline{j}}.$$

7. Если все точки исходной выборки V распределены между классом Ω_j и областью $\Omega_{\overline{j}}$, т.е. $I_1=I$, перейти к п. 8, иначе — вернуться к выполнению п. 4.

8. Провести повторную классификацию точек исходной выборки $V=x^t,\ t=\overline{1,n}$, с помощью решающего правила

$$\overline{m}_{j\overline{j}}(x^t):\begin{cases}x^t\in\Omega_j, \text{ если } \overline{f}_{j\overline{j}}(x^t)>0;\\ x^t\in\Omega_{\overline{j}}, \text{ если } \overline{f}_{j\overline{j}}(x^t)\le 0,\end{cases}$$

где

$$\overline{f}_{j\overline{j}}(x^t)=\frac{1}{n}\sum_{\substack{i=1\\ i\ne t}}^{n}\sigma(i)\prod_{v=1}^{k}\frac{1}{\overline{c}_v}\Phi\left(\frac{x_v^t-x_v^i}{\overline{c}_v}\right).$$

На данном шаге уточняется граница между классом Ω_j и областью $\Omega_{\overline{j}}$, если им соответствуют несимметричные фрагменты оценки плотности вероятности $\overline{p}(x)$. Действия на шаге 8 повторяются до тех пор, пока не будет завершено перераспределение точек между классом Ω_j и областью $\Omega_{\overline{j}}$.

9. Осуществить проверку на однородность класса Ω_j. Для этого в соответствии с пп. 1, 2 проверить возможность разбиения выборки $V_1(j)$ на группы точек, соответствующих одномодальным фрагментам оценки плотности вероятности в области Ω_j. Исследование начинается с уровня p_1 оценки плотности вероятности, при котором ранее были выделены центры класса Ω_j и области $\Omega_{\overline{j}}$. При обнаружении неоднородности выборки $V_1(j)$ осуществляется ее декомпозиция согласно пп. 1—8.

Если в области Ω_j дополнительно классы не выделены, то перейти к обнаружению нового класса в соответствии с приведенной выше методикой, анализируя выборку $V_1(\overline{j})$ в области $\Omega_{\overline{j}}$.

Исследования, результаты которых представлены в настоящей статье, выполнены в рамках Федеральной целевой программы „Научные и научно-педагогические кадры инновационной России“ на 2009—2013 гг., гос. контракт № 02.740.11.0621.

СПИСОК ЛИТЕРАТУРЫ

1. *Дорофеюк А. А.* Алгоритмы автоматической классификации // Автоматика и телемеханика. 1971. № 12. С. 78—113.
2. *Цыпкин Я. З.* Основы теории обучающихся систем. М.: Наука, 1970.
3. Самообучение распознаванию образов по методу смешанных распределений / *В. И. Васильев, В. В. Коноваленко, Ф. П. Овсянникова.* Киев, 1974. (Препринт / АН УССР. Ин-т кибернетики; № 74—30).
4. *Лапко А. В., Лапко В. А., Соколов М. И., Ченцов С. В.* Непараметрические системы классификации. Новосибирск: Наука, 2000.
5. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statistic. 1962. Vol. 33, N 3. P. 1065—1076.
6. *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999.

Сведения об авторах

Александр Васильевич Лапко — д-р техн. наук, профессор; Институт вычислительного моделирования СО РАН, Красноярск; E-mail: lapko@icm.krasn.ru

Василий Александрович Лапко — д-р техн. наук, профессор; Сибирский государственный аэрокосмический университет им. акад. М. Ф. Решетнёва, кафедра космических средств и технологий, Красноярск; E-mail: lapko@icm.krasn.ru

Алексей Николаевич Хлопов — аспирант; Сибирский государственный аэрокосмический университет им. акад. М. Ф. Решетнёва, кафедра космических средств и технологий, Красноярск; E-mail: alexhl@list.ru

Рекомендована СибГАУ

Поступила в редакцию
19.11.10 г.