РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 6 1999

УДК 681.513

И. А. Пестунов
(Красноярск)

БЫСТРЫЕ НЕПАРАМЕТРИЧЕСКИЕ АЛГОРИТМЫ КЛАССИФИКАЦИИ ДЛЯ ОБРАБОТКИ БОЛЬШИХ МАССИВОВ ДАННЫХ

Предлагается простой метод построения непараметрических алгоритмов классификации, быстродействие которых в среднем в десять и более раз превышает быстродействие классификаторов Розенблатта – Парзена, и при этом они лишь незначительно уступают последним по качеству классификации. Приводятся результаты статистического моделирования, подтверждающие эффективность предлагаемого метода.

Введение. В последнее десятилетие значительно возрос интерес к непараметрическим алгоритмам классификации [1–4]. Это связано с тем, что указанные алгоритмы не требуют «жестких» ограничений на вид условных плотностей распределения (унимодальности, нормальности и т. п.) и обеспечивают высокую достоверность распознавания. Однако применение известных (подстановочных) непараметрических алгоритмов классификации при обработке больших массивов данных (например, полученных с помощью датчиков дистанционного зондирования земной поверхности) приводит к неприемлемо большим вычислительным затратам [5]. Поэтому актуальной задачей является разработка вычислительно-эффективных непараметрических алгоритмов распознавания образов.

В настоящей работе представлен простой метод генерации быстрых непараметрических алгоритмов классификации, который основан на сокращении числа представителей обучающей выборки, непосредственно используемых при формировании классификаторов Розенблатта – Парзена.

Метод генерации быстрых непараметрических алгоритмов классификации. Пусть в $X \subseteq \Re^k$ с известными и отличными от нуля вероятностями $q_1, ..., q_M \left(\sum_{i=1}^{M} q_i = 1 \right)$ регистрируются наблюдения из классов $\Omega_1, ..., \Omega_M$ соответственно. Наблюдение из Ω_i есть реализация k-мерного случайного вектора-столбца $x^{(i)} = (x_1^{(i)}, ..., x_k^{(i)})^T \in \Re^k$, плотность распределения которого $f_i(x)$ не известна $(i = \overline{1, M})$, но имеется классифицированная обучающая выборка $V = \bigcup_{i=1}^{M} V^{(i)}$ объемом $N = \sum_{i=1}^{M} N_i$, где $V^{(i)} = \{x_j^{(i)} : x_j^{(i)} \in \Re^k,\ j = \overline{1, N_i}\}$ – множество из N_i независимых наблюдений класса Ω_i. Тогда общий вид ре-

шающего правила (РП) из семейства непараметрических классификаторов Розенблатта – Парзена можно описать формулой

$$\delta_0 = \delta_0(x; V) = \arg\max_{i \in S} \{q_i \hat{f}_{N_i}^{(i)}(x)\}.$$

Здесь $S = \{1, 2, \ldots, M\}$; $\hat{f}_{N_i}^{(i)}(x)$ – непараметрическая оценка плотности $f_i(x)$ в точке $x \in X$, определяемая выражением

$$\hat{f}_{N_i}^{(i)}(x) = \frac{1}{N_i} \sum_{j=1}^{N_i} h^{-k} \Phi\left[\frac{x_1 - x_{j1}^{(i)}}{h}, \ldots, \frac{x_k - x_{jk}^{(i)}}{h}\right],$$

где $\Phi(\cdot)$ – некоторая заданная колоколообразная функция (ядро); h – параметр сглаживания [6].

Вероятность ошибки классификации для РП δ_0 определяется выражением

$$P_N = P_N(\delta_0; V) = \sum_{i=1}^{M} q_i \int_{R^k} \chi_{\hat{X}_i}(x) f_i(x) dx,$$

где $\{\hat{X}_i\}$ – разбиение пространства X, соответствующее РП δ_0; $\chi_{\hat{X}_i}(\cdot)$ – индикаторная функция X_i.

Обратим внимание, что для классификации с помощью правила δ_0 некоторой точки $x \in X$ не требуется знать значения оценок $\{\hat{f}_{N_i}^{(i)}(x)\}$ как таковых, а достаточно лишь решить вопрос о том, какая из величин $q_1 \hat{f}_{N_1}^{(1)}(x), \ldots$ $\ldots, q_M \hat{f}_{N_M}^{(M)}(x)$ больше. На этом основании предлагается в выражении для δ_0 оценки $\{\hat{f}_{N_i}^{(i)}(x)\}$ заменить более простыми (в смысле объема вычислений) статистиками $\{\tilde{f}_{N_i}^{(i)}(x)\}$. Здесь

$$\tilde{f}_{N_i}^{(i)}(x) = \frac{1}{N_i} \sum_{j=1}^{N_i'} h^{-k} \Phi\left[\frac{x_1 - x_{\alpha_i(j),1}^{(i)}}{h}, \ldots, \frac{x_k - x_{\alpha_i(j),k}^{(i)}}{h}\right],$$

где $\alpha_i(j) \in \{1, 2, \ldots, N_i\}$, причем для всех $p \neq q$ $\alpha_i(p) \neq \alpha_i(q)$; $N_i' < N_i$.

В результате такой замены РП δ_0 преобразуется в решающее правило

$$\delta_1 = \delta_1(x; W) = \arg\max_{i \in S} q_i \tilde{f}_{N_i}^{(i)}(x).$$

Здесь $W = \bigcup_{i=1}^{M} W^{(i)}$ есть выборка (назовем ее рабочей) объемом $N' = \sum_{i=1}^{M} N_i'$, где $W^{(i)} = \{x_{\alpha_i(j)}^{(i)} \in V^{(i)},\ j = 1, 2, \ldots, N_i'\}$.

Рабочую выборку W будем формировать из условия сохранения значения оценки вероятности ошибки классификации P_N, полученной методом переклассификации, так как в этом случае гарантируется состоятельность правила δ_1.

Процедура генерации рабочей выборки. Для формирования рабочей выборки W могут быть предложены различные процедуры. Опишем подробно одну из них для двухальтернативного случая.

Обозначим $z_1 = x_1^{(1)}, \ldots, z_{N_1} = x_{N_1}^{(1)}, z_{N_1+1} = x_1^{(2)}, \ldots, z_N = x_{N_2}^{(2)}$. Тогда предлагаемую процедуру можно записать в виде последовательности шагов.

Шаг 1. Классифицировать элементы множества $Z = \{z_i, i = \overline{1, N}\}$ с помощью РП δ_0. Результаты классификации занести в одномерный массив D размерностью N по следующему правилу: $D[i] = 1$, если z_i относится к первому классу, $D[i] = 2$ в противном случае.

Шаг 2. В рабочую выборку W включить точки z_1 и z_{N_1+1}. Из оставшихся точек множества Z сформировать две контрольные выборки $Z1 = \{z_2, z_3, \ldots, z_{N_1}\}$ и $Z2 = \{z_{N_1+2}, z_{N_1+3}, \ldots, z_N\}$. Установить $N' = 2, N'' = 2, p = 0$ и перейти к шагу 5.

Шаг 3. Если $N' = N''$, то процедуру закончить, иначе установить $p = 0$, $N'' = N'$ и перейти к шагу 5.

Шаг 4. Если $p = N$, то перейти к шагу 3.

Шаг 5. Положить $p := p + 1$ и классифицировать точку z_p в соответствии с РП δ_1. Если в результате классификации z_p попадет в первый класс, то перейти к шагу 8.

Шаг 6. Если $D[p] = 2$, то перейти к шагу 4.

Шаг 7. Среди элементов выборки $Z1$ найти точку, ближайшую к z_p, и включить ее в выборку W, исключив при этом из $Z1$. Положить $N' := N' + 1$ и классифицировать точку z_p в соответствии с РП δ_1. Если в результате классификации z_p попадет в первый класс, то перейти к шагу 4, иначе – к началу данного шага.

Шаг 8. Если $D[p] = 1$, то перейти к шагу 4.

Шаг 9. Среди точек выборки $Z2$ найти точку, ближайшую к z_p, и включить ее в выборку W, одновременно исключив из $Z2$. Затем, положив $N' := N' + 1$, классифицировать точку z_p в соответствии с РП δ_1. Если в результате классификации z_p попадет в первый класс, то перейти к началу данного шага, иначе – к шагу 4.

На основе этой процедуры легко построить процедуру формирования выборки W для случая $M > 2$ классов.

Результаты статистического моделирования. Изложенная выше процедура исследовалась на большом количестве как модельных, так и реальных данных. Приведем типичные результаты одного из экспериментов. Рассматривался случай трех равновероятных классов в двумерном признаковом пространстве. Первый класс описывался нормальной плотностью с вектором математического ожидания $\mu^{(1)} = (0,0)^T$ и ковариационной матрицей $\Sigma_1 = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, где параметр σ^2 принимал два значения: $\sigma^2 = 1{,}7$ и $\sigma^2 = 3{,}2$. Для объектов второго класса значения первого признака были распределены равномерно на отрезке [-4, 6], а значения второго признака определялись

в соответствии с выражением $x = -\sqrt{25-(y-1)^2} + \xi$, где y и x есть значения первого и второго признаков соответственно; ξ – случайная величина, имеющая равномерное распределение на отрезке [-1,5, 1,5]. Третий класс генерировался путем зеркального отражения в плоскости XY точек второго класса относительно начала координат. Байесовская вероятность ошибки составляла ~0,03 для $\sigma^2 = 1,7$ и ~0,07 для $\sigma^2 = 3,2$.

Таблица 1

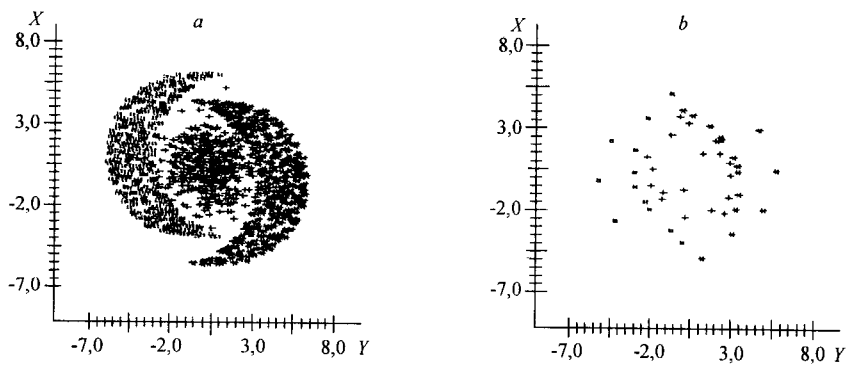
σ^2	$N = N_1 + N_2 + N_3$			
	150	**300**	**600**	**1500**
1,7	**22** 4	**29** 6	**41** 10	**59** 15
3,2	**27** 5	**39** 6	**59** 17	**112** 25

Для данной модели генерировалось по 20 независимых выборок объемом $N = 150, 300, 600$ и 1500 ($N_1 = N_2 = N_3$). Затем строилось РП δ_0 (использовалось нормальное ядро) и для каждой из выборок с помощью метода «скользящего экзамена» определялся наиболее подходящий параметр h. После этого выполнялась процедура генерации рабочих выборок. Полученные результаты представлены в табл. 1 (в каждой ячейке указывается средний объем рабочих выборок (числа выделены) и среднеквадратическое отклонение).

В качестве дополнения к численным результатам на рисунке в графическом виде представлены исходная ($N_1 = N_2 = N_3 = 500$) и рабочая ($N_1' = 12$, $N_2' = 15$, $N_3' = 19$) выборки (для повышения наглядности восприятия обучающих выборок начала координат смещены в отрицательную область).

С целью выяснения вопроса об эффективности предлагаемого метода с точки зрения достоверности распознавания получаемого РП δ_1 было проведено экспериментальное сравнение решающих правил δ_0 и δ_1.

Для данных, относящихся к описанной выше модели, для каждого случая генерировалось по $n = 20$ независимых обучающих выборок объемом $N = 150, 300, 600, 1500$ и по 20 контрольных выборок объемом 1500 (по 500 точек на класс). Затем для полученных обучающих выборок строились РП δ_0 и РП δ_1. Далее для каждой группы из $n = 20$ выборок оценивалось математическое ожидание вероятности ошибки классификации P_N и ее среднеквадратическое отклонение $\hat{\sigma}(P_N)$. Результаты этих вычислений приведены в табл. 2 (значения P_N выделены).



Исходная (*a*) и рабочая (*b*) выборки

Таблица 2

σ^2	РП	$N = N_1 + N_2 + N_3$			
		150	300	600	1500
1,7	δ_0	**0,04** 0,015	**0,031** 0,008	**0,030** 0,006	**0,029** 0,005
	δ_1	**0,062** 0,020	**0,044** 0,012	**0,036** 0,006	**0,030** 0,005
3,2	δ_0	**0,104** 0,022	**0,090** 0,013	**0,083** 0,006	**0,077** 0,005
	δ_1	**0,124** 0,022	**0,099** 0,016	**0,086** 0,008	**0,077** 0,005

Заключение. В данной работе указано на принципиальное отличие задачи оценивания плотности распределения при построении правила классификации от задачи оценивания плотности, рассматриваемой независимо. На основе использования этого факта предложен простой метод генерации вычислительно-эффективных непараметрических алгоритмов классификации для обработки больших массивов данных. Предложенный метод позволяет достичь разумного компромисса между скоростью обработки и достоверностью распознавания.

СПИСОК ЛИТЕРАТУРЫ

1. **Горбачев О. Г., Гребенщиков А. Ю.** Быстрые алгоритмы принятия решения по многомерным данным, использующие обучающие выборки ограниченного объема // Изв. АН СССР. Техн. кибернетика. 1991. № 4. С. 91.
2. **Arinstrong M. L., Abdou I. E.** Nonparametric models and crop classification // Int. Geosci. and Remote Sens. Symp. (IGARSS'85). N. Y. 1985. V. 2. P. 892.
3. **Fukunaga K., Hayes R. R.** The reduced Parzen classifier // IEEE Trans. Pattern Analysis and Machine Intelligence. 1989. **11**, N 4. P. 423.
4. **Харин Ю. С.** Робастность в статистическом распознавании образов. Минск: Университетское изд-во, 1992.
5. **Абрамович Н. С., Ковалев А. А., Плюта В. Е.** К вопросу о классификации природных образований по их оптическим характеристикам в условиях малых выборок // Исследование Земли из космоса. 1985. № 4. С. 105.
6. **Шапиро Е. И.** Непараметрические оценки плотности вероятности в задачах обработки результатов наблюдений // Зарубеж. радиоэлектрон. 1976. № 2. С. 3.

Поступила в редакцию 16 февраля 1999 г.