НЕПАРАМЕТРИЧЕСКАЯ ОЦЕНКА УРАВНЕНИЯ РАЗДЕЛЯЮЩЕЙ ПОВЕРХНОСТИ В УСЛОВИЯХ БОЛЬШИХ ВЫБОРОК И ЕЁ СВОЙСТВА

Лапко А.В., Лапко В.А.
Институт вычислительного моделирования СО РАН, Россия, lapko@icm.krasn.ru

Вычислительная эффективность непараметрических алгоритмов распознавания образов, которые основаны на оценках плотности вероятности типа Розенблатта-Парзена [1], снижается в условиях обучающих выборок большого объёма.

В работе [2] предложена методика синтеза двухуровневых непараметрических систем классификации, обеспечивающих использование технологии параллельных вычислений. Идея подхода состоит в декомпозиции обучающей выборки по её объёму, построении на этой основе семейства непараметрических уравнений разделяющей поверхности между классами и их обобщении в коллективе решающих функций. Исследованы асимптотические свойства подобного коллектива непараметрических решающих функций в двуальтернативной задаче распознавания образов для одномерного случая с учётом неравномерности распределения элементов обучающей выборки между классами [3].

В данной работе полученные результаты обобщаются на многомерную задачу классификации и устанавливаются количественные зависимости аппроксимационных свойств коллектива непараметрических решающих функций от параметров структуры системы и особенностей обучающей выборки.

Коллектив многомерных непараметрических уравнений разделяющей поверхности. Пусть $V = \left(x^i, \sigma\left(x^i\right),\ i = \overline{1,n}\right)$ - обучающая выборка объёма n, составленная из признаков $x^i = \left(x_j^i,\ j = \overline{1,k}\right)$ классифицируемых объектов и соответствующих «указаний учителя» $\sigma\left(x^i\right)$ об их принадлежности к одному из двух классов Ω_1, Ω_2:

$$\sigma\left(x^i\right) = \begin{cases} -1, \textit{если } x^i \in \Omega_1 \\ 1, \textit{если } x^i \in \Omega_2 \end{cases}.$$

Объём n обучающей выборки V является большим, что снижает вычислительную эффективность непараметрических алгоритмов распознавания образов.

В соответствии с методикой, изложенной в работе [2], осуществим декомпозицию исходной выборки V на части $V_j = \left(x^i, \sigma\left(x^i\right),\ i \in I_j\right)$, $j = \overline{1,N}$, где I_j - множество номеров ситуаций из V, составляющих j-ю группу. Количество элементов n_j множества I_j, $j = \overline{1,N}$ одинаково и равно $\bar{n}$, причём $N = n/\bar{n}$.

По полученным данным V_j, $j = \overline{1,N}$ построим семейство непараметрических оценок уравнения разделяющей поверхности

$$\bar{f}_{12}^{\,j}(x) = \bar{p}_2^{\,j}(x) - \bar{p}_1^{\,j}(x),\ j = \overline{1,N}, \qquad (2)$$

где непараметрическая оценка плотности вероятности распределения признаков x анализируемых объектов в s-м классе представляются статистикой типа [1]

$$\bar{p}_s^{\,j}(x) = \frac{1}{\bar{n}_s \prod\limits_{v=1}^{k} c_v^j} \sum_{i \in I_j} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v^j}\right),\ s = 1,2. \qquad (3)$$

В выражении (3) ядерные функции $\Phi(u_v)$ удовлетворяют условиям H:

$$\Phi(u_v)=\Phi(-u_v),\quad 0\le\Phi(u_v)<\infty,$$

$$\int\Phi(u_v)du_v=1,\quad \int u_v^2\,\Phi(u_v)du_v=1,$$

$$\int u_v^m\,\Phi(u_v)du_v<\infty,\quad 0\le m<\infty;\ v=\overline{1,k},$$

$c_v, v=\overline{1,k}$ - коэффициенты размытости ядерных функций, значения которых убывают с ростом количества n_j элементов множества I_j, $j=1,N$.

Здесь и далее бесконечные пределы интегрирования опускаются.

В качестве обобщённой решающей функции в двуальтернативной задаче распознавания образов примем статистику вида

$$\bar{\bar{f}}_{12}(x)=\frac{1}{N}\sum_{j=1}^{N}\bar{f}_{12}^{\,j}(x). \tag{4}$$

Оптимизация непараметрического решающего правила

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

по коэффициентам размытости ядерных функций c_v, $v=\overline{1,k}$ осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов.

Асимптотические свойства коллектива непараметрических решающих функций. Для получения аналитически значимых результатов будем считать, что интервалы изменения признаков x_v, $v=\overline{1,k}$ классифицируемых объектов одинаковы. В этих условиях появляется возможность полагать, что коэффициенты размытости c_v, $v=\overline{1,k}$ ядерных функций соизмеримы, т.е. $c_v=c$, $v=\overline{1,k}$.

Теорема. Пусть плотности вероятности $p_j(x)$, $j=1,2$ распределения многомерной случайной величины $x=(x_1,\ldots,x_k)$ в классах и первые две их производные $p_{jv}^{(1)}(x)$, $p_{jv}^{(2)}(x)$, $j=1,2$ по каждой компоненте x_v, $v=\overline{1,k}$ ограничены и непрерывны; ядерные функции $\Phi(u)$ удовлетворяют условиям нормированности, положительности и симметричности H; последовательности $c(\bar{n})=c$ коэффициентов размытости ядерных функций в статистиках $\bar{f}_{12}^{\,j}(x)$, $j=1,N$ таковы, что при $\bar{n}_1\to\infty$, $\bar{n}_2\to\infty$, значения $c\to 0$, а $\dfrac{\bar{n}_1+\bar{n}_2}{\bar{n}_1\,\bar{n}_2\,c^k}\to 0$. Тогда при конечных значениях N многомерная непараметрическая оценка $\bar{\bar{f}}_{12}(x)$ байесовского уравнения разделяющей поверхности $f_{12}(x)=p_2(x)-p_1(x)$ обладает свойством асимптотической несмещенности

$$M\left(f_{12}(x)-\bar{\bar{f}}_{12}(x)\right)\sim\frac{c^2}{2}\sum_{v=1}^{k}\left(p_{2v}^{(2)}(x)-p_{1v}^{(2)}(x)\right) \tag{6}$$

и сходимости в среднеквадратическом

$$M\int\ldots\int\left(f_{12}(x)-\bar{\bar{f}}_{12}(x)\right)^2dx_1\ldots dx_k\sim$$

$$\sim\frac{(\bar{n}_1+\bar{n}_2)\prod_{v=1}^{k}\int\Phi^2(u_v)du_v}{N\,\bar{n}_1\,\bar{n}_2\,c^k}+\frac{c^4}{4}B. \tag{7}$$

Здесь M - знак математического ожидания, а значение определяется выражением

$$B = \int \ldots \int \left(\sum_{v=1}^{k} \left(p_{2v}^{(2)}(x) - p_{1v}^{(2)}(x) \right) \right)^2 dx_1 \ldots dx_k .$$

Из свойств (6), (7) следует свойство состоятельности статистики $\bar{\bar{f}}_{12}(x)$.

Доказательство теоремы основывается на технологии исследования асимптотических свойств коллектива непараметрических уравнений разделяющей поверхности, приведённых в работе [3] для одномерного случая.

При оценивании уравнения разделяющей поверхности $f_{12}(x)$ по выборке V без предварительной ее декомпозиции ($N = 1$), полученные результаты (7) совпадают с асимптотическими свойствами традиционной непараметрической оценки решающей функции парзеновского типа.

Сравнение асимптотических свойств непараметрических оценок уравнений разделяющих поверхностей. Исследуем аппроксимационные свойства статистики (4) по сравнению с традиционной непараметрической оценкой уравнения разделяющей поверхности

$$\tilde{f}_{12}(x) = \frac{1}{n c^k} \sum_{i=1}^{n} \sigma\left(x^i\right) \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c} \right), \tag{8}$$

Восстанавливаемой по исходной выборки V в условиях неравномерности распределения её элементов между классами.

Будем считать, что $n_1 = \alpha n$, $n_2 = (1-\alpha)n$, $\alpha \in (0;1)$. Причём синтез составляющих непараметрической оценки решающей функции (4) осуществляется на основе выборки объёма $\bar{n} = \frac{2\alpha n}{N}$ при равномерном распределении элементов обучающих выборок между классами, т.е. $\bar{n}_1 = \bar{n}_2 = \frac{\alpha n}{N}$. В этих условиях синтез статистики $\bar{\bar{f}}_{12}(x)$ осуществляется по выборке меньшего объёма по сравнению с $\tilde{f}_{12}(x)$.

В качестве критерия эффективности используются отношения R_j, $j = \overline{1,3}$ асимптотических выражений смещения, среднеквадратического отклонения статистик $\bar{\bar{f}}_{12}(x)$, $\tilde{f}_{12}(x)$ от байесовской решающей функции $f_{12}(x)$ и их дисперсий при оптимальных значениях коэффициентов размытости ядерных функций.

В принятых выше условиях выражение (7) представляется в виде

$$\frac{2N \prod_{v=1}^{k} \int \Phi^2(u_v)\, du_v}{\alpha n c^k} + \frac{c^4}{4} B, \tag{9}$$

а соответствующее асимптотическое выражение среднеквадратического отклонения традиционной непараметрической оценки уравнения разделяющей поверхности (8) от байесовской решающей функции запишется как

$$\frac{\prod_{v=1}^{k} \int \Phi^2(u_v)\, du_v}{\alpha(1-\alpha) n c^k} + \frac{c^4}{4} B. \tag{10}$$

Определим минимальное значение W_2^r, $W_2(\alpha)$ выражений (9), (10) при оптимальных значениях соответствующих им коэффициентов размытости ядерных функций

$$c^* = \left(\frac{2k\,N \prod_{v=1}^{k} \int \Phi^2(u_v)\,du_v}{\alpha\, n\, B} \right)^{1/(k+4)}, \quad \bar{c} = \left(\frac{k \prod_{v=1}^{k} \int \Phi^2(u_v)\,du_v}{\alpha(1-\alpha) n\, B} \right)^{1/(k+4)}.$$

Тогда, после несложных преобразований получим

$$R_2 = \frac{W_2^r}{W_2(\alpha)} = \left[\frac{(2(1-\alpha))^4}{N^k} \right]^{1/k+4} \frac{4 + N\,k}{4 + k^{k/(k+4)}}.$$

Заметим, что первые слагаемые $W_3^r(c)$, $W_3(\alpha,c)$ выражений (9), (10) соответствуют главным дисперсионным составляющим статистик $\bar{\bar{f}}_{12}(x)$, $\tilde{f}_{12}(x)$, вторые – квадрату их смещения $W_1^r(c)$, $W_1(\alpha,c)$.

Следуя принятой технологии исследования, вычислим отношения

$$R_3 = \frac{W_3^r(c^*)}{W_3(\alpha,\bar{c})} = \left[\frac{(2(1-\alpha))^4}{N^k} \right]^{\frac{1}{k+4}},$$

$$R_1 = \frac{W_1^r(c^*)}{W_1'(\alpha,\bar{c})} = (2(1-\alpha)N)^{2/(k+4)}$$

при оптимальных значениях коэффициентов размытости c^*, $\bar{c}$ статистик $\bar{\bar{f}}_{12}(x)$, $\tilde{f}_{12}(x)$.

При $k=1$ полученные отношения совпадают с результатами работы [3], что подтверждает корректность выполненных преобразований.

Анализ полученных результатов показывает существование пороговых значений $\bar{\alpha}$ степени неравномерности распределения α, при которых сохраняются соотношения $R_3 < 1$. С ростом N значения $\bar{\alpha}$ имеют тенденцию к снижению.

С увеличение количества N частей обучающей выборки, на которые она разбивается в процессе синтеза коллектива непараметрических решающих функций, при малых значениях k наблюдается существенный рост значений R_1, R_2 по сравнению с темпом снижения значений R_3. С уменьшением α значения исследуемых отношений возрастают при сохранении тенденции их изменения. С увеличение k значения R_1, R_2 асимптотически стремятся к 1.

Предложенный коллектив непараметрических оценок решающих функций в двуальтернативной задаче распознавания образов, основанный на декомпозиции обучающей выборки по её объёму, обеспечивает использование технологии параллельных вычислений.

Литература

1. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statistic, 1962, Vol.33. – p. 1065-1076.
2. Lapko A.V., Lapko V.A. Nonparametric Pattern Recognition Systems Based on Learning a Sample Decomposition by Its Dimension // Pattern recognition and image analysis, 2009. – Vol. 19. - №2. – P. 296 - 302.
3. Лапко А.В., Лапко В.А. Коллектив непараметрических решающих функций в двуальтернативной задаче распознавания образов // Системы управления и информационные технологии, 2009. – 3.1(37). – С. 156 – 160.