第 14 卷 第 2 期
1988 年 3 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 14, No. 2
Mar., 1988

补缀式浓缩近邻分类器 BDPATCH

王 庆 人
(南 开 大 学)

摘 要

本文提出一种浓缩近邻分类器 BDPATCH. 其浓缩集从编辑过的训练集经显露和补缀边界模式产生，具有 Bayes 渐近最优性. 对 BDPATCH 和其它已有的 CNN 算法进行了比较，结果表明这种新的分类器具有高识别率，同时又是快速的.

一、导 言

文献[1, 2]已经证明，k-近邻 (k-NN) 分类方法在 $n \to \infty$，$k \to \infty$ 和 $k/n \to 0$ 时是渐近 Bayes 最优的，n 为训练样本容量. 但在 n 和 k 很大时占用内存和消耗机时很多；在分布函数复杂时，k 的最佳值也难以估计. 浓缩近邻 (CNN) 分类方法[3]在时间、空间上的效率都有较大提高. 这种方法从训练样本中抽取一个浓缩子集，使得：1) 全部训练模式可以被浓缩子集和 1-近邻法正确分类；2) 浓缩集越小越好. 文献 [4—7] 对于 CNN 法提出的各种改进都以浓缩集的最小化为目标. 事实上仅仅给训练模式正确分类并不保证高识别率. 目前尚无关于 CNN 法识别率的分析研究或渐近最优性的报道.

编辑近邻法 (ENN)[8] 是 k-NN 的另一种改进：对训练样本施行编辑后，属于不同类别的模式不再交叉分布. ENN 的识别率已被证明是依概率渐近 Bayes 最优的[9].

本文提出一种高性能的浓缩近邻分类器 BDPATCH，它包括编辑、显露和补缀三个阶段. 该法同时考虑了识别率和效率两方面的问题.

二、BDPATCH 浓缩过程及渐近最优性

1. Voronoi 多面体

设给定的识别问题有 m 个类别，每个模式表示为距离空间 D 的一个向量，S_n 为容量是 n 的训练样本

$$S_n = \{(X_i, \Theta_i) | X_i \in D, \Theta_i \in \{1, 2, \cdots, m\}, i = 1, 2, \cdots, n\}. \quad (1)$$

其中 Θ_i 是模式 (X_i, Θ_i) 的类别. 对 S_n 施行编辑后得到 $S_n^E \subset S_n$. 在 S_n^E 中可以按文献[10]的方法对每一个模式 (Y, Θ) 定义一个 Voronoi 多面体

本文于 1985 年 10 月 29 日收到.

$$V(Y, \Theta) = \{X \in D \mid \forall (Y', \Theta') \in S_n^E, \ |X, Y| \leqslant |X, Y'|\}. \tag{2}$$

其中 $|X, Y|$ 表示X与Y之间的距离；$V(Y, \Theta)$ 实际上由与Y最近的那些点组成。图1表示了各个 Voronoi 多面体，它们共同构成 S_n^E 的 Voronoi 图。ENN 的 m 个分类区域实际上由各类模式的 Voronoi 多面体并成，即

$$V_i^n = \bigcup_{\Theta = i} V(Y, \Theta),$$

$$i = 1, 2, \cdots, m. \tag{3}$$

因此 ENN 的识别率可以表示为

$$R_n^E = \sum_{i=1}^{m} \pi_i P_i(V_i^n). \tag{4}$$

其中 π_i 为第 i 类模式的先验概率；P_i 为其概率测度。根据文献[9]，当合理地控制编辑过程时，例如象轮番编辑 (MULTIEDIT) 或半保留法 (Holdout) 那样，R_n^E 依概率趋近于 Bayes 识别率 R^*.

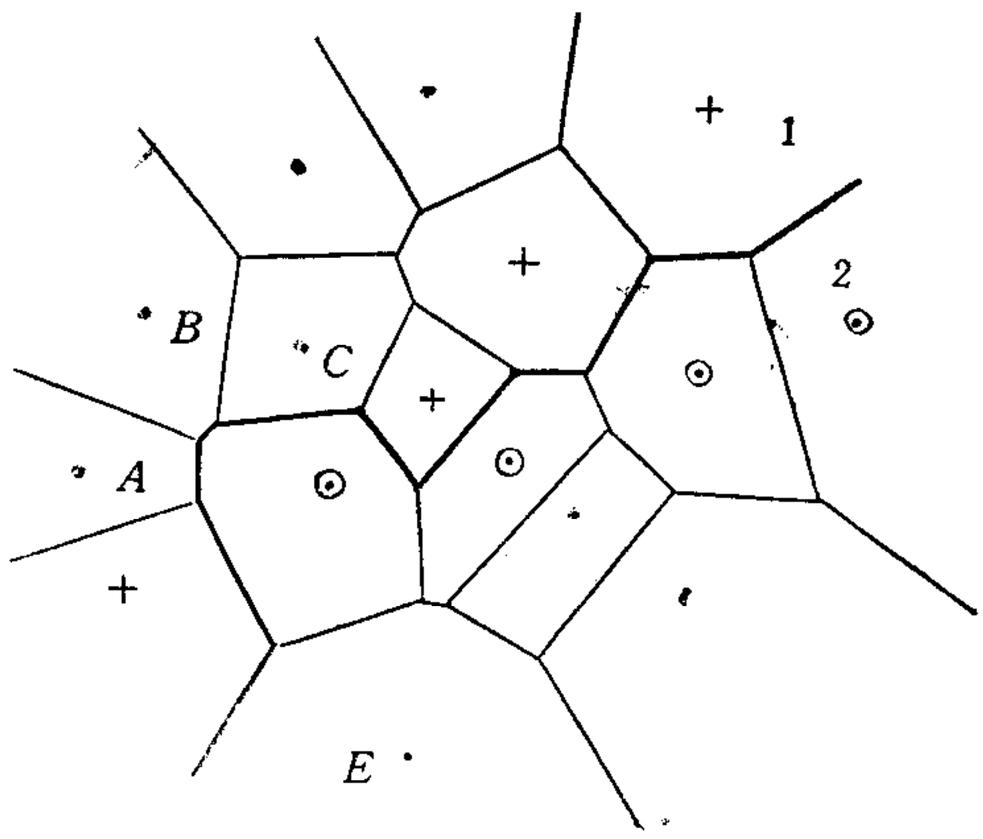

图1　S_n^E 的 Voronoi 图
EBD = BDU{A, B, C, E}

2. 边界显露过程

如果一个 Voronoi 多面体与某个 V_i^n 有公共面，我们就称它的相应模式为边界模式。设 EBD 为 S_n^E 中全体边界模式构成的集合。可以证明，如果以 EBD 为浓缩集，则 CNN 的识别率也为 R_n^E. 下面定义子集合 B_i, $i = 1, 2, \cdots, m$.

$$B_i = \bigcup_{(X, i) \in S_n^E} \{(Y, \Theta) \in S_n^E, \Theta \neq i \mid \forall (Y', \Theta') \in S_n^E, \Theta' \neq i \rightarrow |X, Y| \leqslant |X, Y'|\} \tag{5}$$

很容易证明

$$B_i \subset \text{EBD}, \ i = 1, 2, \cdots, m. \tag{6}$$

B_i 实际由非 i 类模式组成，并且每个这样的模式都是全体非 i 类模式中距离某个 i 类模式最近的。再定义

$$\text{BD} = \bigcup_{i=1}^{m} B_i. \tag{7}$$

一般地 BD ≠ EBD，但包含了 EBD 中相当数量甚至大部分模式，并且容易根据定义(5),(7)求取。这就是下面的显露程序过程 EXPOSING.

```
Procedure EXPOSING;
    For i←——1 to m do
    Begin
      1) For each (X, i) ∈ S_n^E do
           find (Y, Θ) ∈ S_n^E Such that Θ≠i and
             |X, Y| ≤ |X, Y'| for (Y', Θ') ∈ S_n^E, Θ' ≠ i
      2) Collect (Y, Θ) found at step 1 to form B_i
```

End;

$$BD = \bigcup_{i=1}^{m} B_i$$

End EXPOSING;

从图 1 可以看出，BD 包括了 EBD 的大多数模式，但不是全部的，因而若以 BD 为浓缩集，CNN 的识别率与 R_n^E 差别还比较大. 例如图 1 中的 BD 就不包括 EBD 中的 A, B, C 和 E 点.

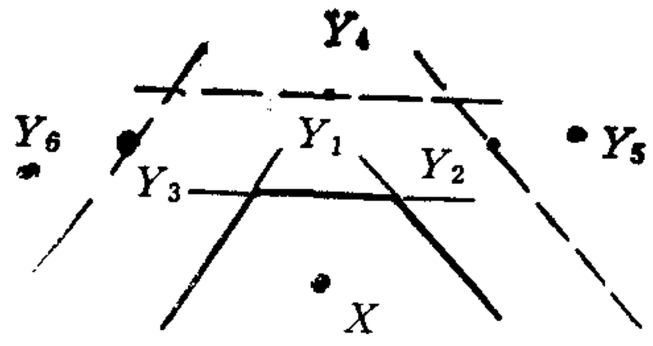

图 2 在X点处的 PATCHING

3. 边界补缀过程

我们可以对 BD "补缀"，使之扩大到十分接近EBD. 补缀过程从 B_i 的任一模式开始，见图 2. 设 X 是位于边界的一个 i 类模式，离它最近的非 i 类模式 $Y_1 \in B_i$. 补缀从 Y_1 开始. 先把它放入 PATCH 集合. 过 Y_1 做平行于 X 与 Y_1 之间界面的平面，凡在此平面的远离 X 一侧的模式都几乎不可能属于 EBD，因而暂排除考虑. 例如 Y_4 被排除考虑. 在其余的非 i 类模式中，Y_2 距 X 最近，因而它必属于 EBD，被补入 PATCH. 同理过 Y_2 做平行于边界的平面后，Y_5 被排除考虑；Y_3 被补入 PATCH 后，Y_6 又被排除. 这个过程迭代地进行，直到每个 B_i 都不再扩大为止. 下面是用类 Pascal 语句描述的补缀 (PATCHING) 程序过程. 这里假定 B_i 包含 n_i 个模式，ST_i, i, j, k 和 l 为局部整数变量，X, Y, Z 和 W 为局部向量变量，PATCH 为局部向量集合变量.

Procedure PATCHING

$ST_i \longleftarrow 1$ for $i = 1, 2, \cdots, m$

Repeat

for $i \longleftarrow 1$ to m do

Begin

for $j \longleftarrow ST_i$ to n_i do

Begin

$X \longleftarrow$ the jth pattern in B_i;

$(Y, \Theta) \longleftarrow$ the closest pattern to X in $S_n^E - T_n^i$

where $T_n^i = \{(W, \Theta) \in S_n^E \mid \Theta = i\}$;

PATCH $\longleftarrow \{(Y, \Theta)\}$;

every pattern in $S_n^E - T_n^i$ is made effective;

$k \longleftarrow 1$;

Repeat

let (Y, Θ) be the kth pattern in PATCH;

$Z \longleftarrow (Y - X)/|Y - X|$;

for each $W \in S_n^E - T_n^i$ do

if $Z \cdot (W - X) \geqslant 0$

then make W ineffective;

find the closest effective pattern in $S_n^E - T_n^i$;

if it ever exists, let it be assigned to (Y, Θ)

```
            PATCH←——PATCH + {(Y, Θ)}
            k←——k + 1
        Until PATCH remains unchanged;
        for k←——1 to m, k⟨⟩i do
          B_k←——B_k + PATCH ∩ T_n^k;
        increment n_k
      End
  End
Until every B_i (i = 1, 2, ···, m) remains unchanged
End PATCHING;
```

4. 渐近最优性

图 3 解释了 BD 经补缀后，加进了模式 A，C 和 E，但还比 EBD 少一个模式 B，从 A 开始或从 C 开始补缀时 B 被排除了．因为 B 位于过 A 的和过 C 的那种平行于界面的平面之外（图 2）．总之，从邻近 B 的每一个模式看去，B 都是远离边界的．因此 B 的被排除，在 ENN 法与用 BD 为浓缩集的 CNN 法间差别不大．其差别在图 3 中只局限于标为 Δ 的很小区域．如果记 CNN 的识别率为 R_n^C，则

$$|R_n^C - R_n^E| \leqslant P(\Delta). \qquad (8)$$

而 $P(\Delta)$ 是微乎其微的．既然已知 R_n^E 依概率趋近 R^*，依式 (6)，(8) 可以期望 R_n^C 依概率趋近 R^*．因此可以得到直观的结论：Editing—EXPOSING—PATCHING 产生的浓缩集提供了渐近 Bayes 最优的 CNN 法．

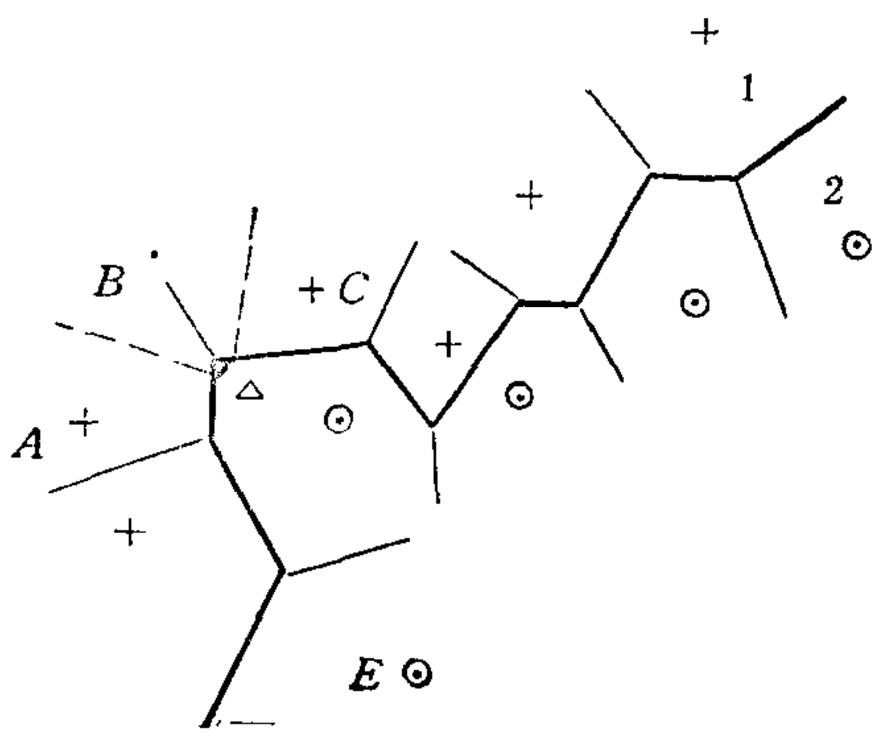

图 3　经 PATCHING 后的 BD EBD = BD∪{B}

三、仿真试验和结论

我们在 Fujitsu M-340S 计算机上对 BDPATCH 和其它六种 CNN 法进行了仿真试验．训练样本与检验样本独立地按下述两类问题的分布密度生成，先验概率 $\pi_1 = \pi_2 = 1/2$：

$$f_1(x, y, z) = \frac{1}{2 \cdot (1.15\sqrt{2\pi})^3} \exp\left[-\frac{(x-2)^2 + y^2 + (z-1)^2}{2 \cdot 1.15^2}\right] + \frac{1}{2 \cdot (1.05\sqrt{2\pi})^3} \exp\left[-\frac{(x+2)^2 + y^2 + (z-1)^2}{2 \cdot 1.05^2}\right], \qquad (9)$$

$$f_2(x, y, z) = \frac{1}{2 \cdot (0.75\sqrt{2\pi})^3} \exp\left[-\frac{x^2 + (y-2)^2 + (z+1)^2}{2 \cdot 0.75^2}\right] + \frac{1}{2 \cdot (0.65\sqrt{2\pi})^3} \exp\left[-\frac{x^2 + (y+2)^2 + (z+1)^2}{2 \cdot 0.65^2}\right]. \qquad (10)$$

这个识别问题的分类边界是复杂的.其 Bayes 识别率用 Monte-Carlo 法估计为 0.95295，或误识率为 0.04750.

被仿真的近邻分类器包括

1) HART-68: 最原始的 CNN[3].

2) EBHART: 连续施用 MULTIEDIT[9], EXPOSING 和 HART-68.

3) REDUCE: 施用 HART-68 后再约简浓缩集[4].

4) EBREDU: 连续施用 MULTIEDIT, EXPOSING 和 REDUCE.

5) SELTV: 是文献[5]所提出的 SELECTIVE 的一个简化,即将其最后一步的动态规划改为启发式搜索.

6) EBSELT: 连续施用 MULTIEDIT, EXPOSING 和 SELTV.

7) BDPATCH: 本文所提浓缩算法，即连续施用 MULTIEDIT, EXPOSING 和 PATCHING.

试验共使用了 15 个训练集，其中每三个容量相同的为一组. 这五个容量值分别为 $n = 48, 96, 192, 400$ 和 800. 对每一容量值下每一 CNN 法进行三次仿真. 用独立生成的容量为 1200 的检验样本估计每一情况下的识别率、分类速度、内存占用和设计机时. 仿真结果列于表 1. 为了便于比较,把编辑近邻法 MULTIEDIT 的性能列于标为 MTEDIT 的一列中.

表 1 CNN 分类器平均性能

项 目	n	HART-68	EBHART	REDUCE	EB-REDU	SELTV	EB-SELT	BDPATCH	MTEDIT
训练机时(秒)	48	0.13	0.65	0.79	0.92	0.55	0.38	0.74	0.68
	192	1.51	5.89	18.68	14.11	9.57	8.21	6.90	4.70
	800	23.22	74.58	---	293.61	202.71	173.08	82.55	32.30
内存开销	48	15.33	11.00	12.00	6.00	12.67	5.67	19.00	42.33
	192	34.67	32.00	28.33	12.67	29.00	13.00	58.33	168.33
	800	115.00	60.67	---	34.00	88.33	31.67	118.00	726.33
分类机时(毫秒)	48	1.33	0.96	1.05	0.54	1.11	0.52	1.64	3.68
	192	2.97	2.69	2.43	1.08	2.49	1.13	4.87	15.48
	800	9.82	5.06	---	2.86	7.54	2.72	10.04	62.99
误识率	48	0.1167	0.1175	0.1278	0.1067	0.1261	0.0972	0.0891	0.0856
	192	0.0847	0.0744	0.0939	0.0775	0.1019	0.0636	0.0567	0.0564
	800	0.0786	0.0622	---	0.0617	0.0903	0.0644	0.0522	0.0517

观察和分析试验结果,得到如下结论:

(1) 如文献[4,5]所期望的，REDUCE 和 SELTV 比 HART-68 的浓缩率要高,因此分类效率也高. 但识别率有所下降. 原版 SELECTIVE[5] 的识别率高于 SELTV，但未考虑实际识别过程的误识率. 按照 SELTV 与 HART-68 的关系类推，SELECTIVE 的识别率可能更低. 上述比较同样适用于 EBHART, EBREDU 和 EBSELT 之间. 根据表 1 的数字,增大浓缩率得到的效率增益并不明显,但识别率的下降是明显的. 因此在研究和比较 CNN 方法时,必须考虑实际识别过程的识别率. 本文是同类研究中第一次这样做的.

(2) 如同文献[9]和本文所期望的,先编辑再浓缩比单纯浓缩效果要好得多. 本文的

前处理除包括编辑外，还用 EXPOSING 求取 BD 集．BD 中的模式全部是边界模式，从它出发的补缀是可靠的．

(3) MULTIEDIT 表现出渐近 Bayes 最优性．我们新发展的 BDPATCH 具有高的识别率．两种分类器的识别率几乎没有什么差别，因此 BDPATCH 也表现出渐近最优性．但 BDPATCH 的时间、空间效率高得多．例如当 $n = 800$ 时，BDPATCH 快 6 倍，而且倍数是随着 n 的增大而上升的．

(4) BDPATCH 的分类速度比其它 CNN 法低些，但识别率明显高．而且是几种 CNN 算法中，唯一表现出象 MULTIEDIT 那样的渐近最优性的．此外 BDPATCH 所用的训练机时少于或等于其它 CNN 法．

参　考　文　献

[1] Cover, T. M. and Hart, P. E., Nearest Neighbor Pattern Classification, *IEEE Trans. Inform. Theory,* IT-**13**(1967), 21—27.

[2] Devroye, L., Some Property of the K-nearest Neighbor Rule, Proc. 5th Int. Conf. Pattern Recognition (1980), 103—105.

[3] Hart, P. E., The Condensed Nearest Neighbor Rule, *IEEE Trans. Inform. Theory,* IT-**14**(1968), 515—516.

[4] Gates, G. W., The Reduced Nearest Neighbor Rule, *IEEE Trans. Inform. Theory,* IT-**18**(1972), 431—433.

[5] Ritter, G. L., Woodruff, H. B., Lowry, S. R. and Isenbour, T. L., An Algorithm for a Selective Nearest Neighbor Decision Rule, *IEEE Trans. Inform. Theory,* IT-**21**(1975), 665—669.

[6] Tomek, I., Two Modifications of CNN, *IEEE Trans. Syst., Man, Cybern.,* SMC-**6**(1976), 769—772.

[7] Gowda, K. C., and Krishna G., The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighborhood, *IEEE Trans. Inform. Theory,* IT-**25**(1979), 488—490.

[8] Wilson, D. L., Asymptotic Property of NN-rules Using Edited Data, *IEEE Trans. Syst., Man. Cybern.,* SMC-**2**(1972), 408—421.

[9] Devijver, P. A. and Kittler, J., Pattern Recgnition, A Statistical Approach, Prentice-Hall Int., Inc. (1982).

[10] Toussaint, G. T., Pattern Recognition and Geometrical Complexity, Proc. 5th Int. Conf. Pattern Recognition (1980), 1324—1327.

A CNN CLASSIFICATION DESIGN WITH BOUNDARY PATCHING

WANG QINGREN

(*Nankai University*)

ABSTRACT

A condensed nearest neighbor classification rule, BDPATCH, is proposed. The condensed set is produced by exposing and patching the boundary pattern subset of the edited training set. This procedure results in a Bayes asymptotically optimal classifier. Simulation experiment on this new and other existing CNN rules is presented, which shows that BDPATCH can achieve a high recognition rate and maintain a fast speed.