Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 121-126, April 02-05, 2001 - PUC, Rio de Janeiro - RJ - Brazil

Complexidade de Amostra para Projetar Operadores para Imagens Binárias pela Aprendizagem de Máquina

Hae Yong Kim
Dept. Eng. de Sistemas Eletrônicos, Escola Politécnica, Universidade de São Paulo
Av. Prof. Luciano Gualberto, trav. 3, 158; CEP 05508-900, São Paulo, SP, Brasil
hae@lps.usp.br, http://www.lps.usp.br/~hae

Abstract

Binary operators (or filters) have a broad range of pratical applications. Several works have shown that binary operators can be sucessfully designed using a system based on the machine learning. Designing a binary operator by hand is a hard and annoying task. The use of a learning system allows the user to specify operators easily, by simply feeding the system with pairs of in-out sample images. In this paper, we present a set of techniques to estimate the sample complexity of the binary operator learning problem. The sample complexity is the quantity of training samples needed to get, with probability at least (1-δ), an operator with an error rate at most ε. We make use of the PAC (Probably Approximately Correct) learning theory to compute it, to both noise-free and noisy cases. As the PAC theory usually overestimates the sample complexity, the statistical estimation is used to calculate *a posteriori*, a tight error rate. We also show how the minimal error rate for a given noisy problem can be estimated. Finally, we apply the theory developed to analyze the sample complexity and the error rate for the spatial resolution increasing of electronic documents with printed characters by machine learning.

1. Introdução

O processamento de imagens binárias é um ramo importante do processamento de imagens. Para justificar essa afirmação, podemos citar, entre muitas outras razões:

- A maioria dos documentos eletrônicos são imagens binárias.
- A binarização muitas vezes é a primeira etapa da visão computacional ou de OCR e todo o resto do processamento pode ocorrer inteiramente sobre imagens binárias.
- A grande maioria das impressoras de jato de tinta ou laser atualmente utilizadas não são capazes de imprimir em tons de cinza verdadeiros: imprimem somente minúsculos pontos pretos. Assim, qualquer imagem em níveis de cinza deve ser convertida numa imagem binária por um processo chamado halftone antes de ser enviado à impressora.

Assim, as técnicas para se projetar um operador (às vezes também chamado de filtro) binário apropriado para uma determinada tarefa são importantes, pois o seu projeto manual é normalmente difícil e tedioso. Diversos trabalhos têm mostrado que os operadores em geral (binários, níveis de cinza, coloridos, multi-espectrais, etc.) podem ser projetados com sucesso através de sistemas baseados em aprendizagem de máquina [1, 2, 3]. Porém a aprendizagem de máquina é especialmente apropriada para se projetar os operadores binários, pois neste caso é possível projetar operadores definidos em janelas relativamente grandes com um custo computacional tolerável. Por exemplo, [4] utiliza esta idéia para reconhecer caracteres impressos sem segmentação, [5] utiliza-a para eliminar ruídos, enquanto que [6, 7] utilizam-na para aumentar a resolução de documentos binários.

Um operador definido numa janela (denotado W-operador) é uma transformação de imagem onde a cor de um pixel de saída depende unicamente das cores da sua vizinhança na imagem de entrada. Estamos interessados em projetar este tipo de operador.

Para se usar um sistema baseado em aprendizagem, é necessário ter meios para estimar a complexidade de amostra. A complexidade de amostra é a quantidade de amostras de treinamento necessários para se obter, com probabilidade maior que (1-δ), um operador com uma taxa de erro menor que ε. Com alguma experiência, o usuário consegue estimar empiricamente o tamanho das imagens de treinamento necessárias para se obter uma qualidade "razoável", porém evidentemente isto não é nada científico.

Assim, o presente trabalho pretende estudar as diferentes técnicas para se determinar a complexidade de amostra e a taxa de erro. Muitos trabalhos anteriores [5, 1, 6] ou supõem que a verdadeira distribuição de probabilidade é conhecida ou então (o que no fundo é equivalente) supõem que a estatística obtida a partir das imagens amostras é uma aproximação muito boa da distribuição de probabilidade verdadeira. Ora, na prática quase nunca a verdadeira distribuição é conhecida. Também não é lícito supor que a estatística das imagens seja uma boa aproximação da distribuição verdadeira, pois basta que as imagens de treinamento não sejam suficientemente grandes para que a aproximação não seja boa.

Para atingir o nosso objetivo primeiro utilizaremos a teoria de aprendizagem PAC (Provavelmente Aproximadamente Correta), analisando os casos sem ruído e ruidoso. Como a teoria PAC costuma superestimar a complexidade de amostra, descreveremos as técnicas de estimação estatística para estimar, depois de se ter projetado o operador, qual foi a verdadeira taxa de erro obtida. Além disso, no caso ruidoso, existe uma taxa de erro mínima, sendo impossível projetar um operador com uma taxa de erro menor que esta mínima. Assim, também descrevemos como é possível estimar o erro mínimo. Por fim, aplicamos a teoria desenvolvida para analisar a complexidade de amostra e a taxa de erro para o problema de aumento de resolução espacial de documentos com caracteres impressos.

2. O Problema

Faremos uso da teoria de aprendizagem PAC (Provavelmente Aproximadamente Correta) para calcular a complexidade de amostra do problema de aprendizagem de operadores binários para imagens. Infelizmente, com freqüência, somente uma complexidade de amostra superestimada pode ser obtida utilizando esta teoria. Mesmo assim, ela será útil como um limite superior para a quantidade de amostras necessárias, e para mostrar a convergência do processo de

aprendizagem. Além disso, a teoria de aprendizagem PAC irá nos permitir expressar rigorosamente o problema de aprendizagem do W-operador, e pode clarificar consideravelmente a compreensão do problema.

Vamos definir uma imagem binária como uma função $Q:\mathbb{Z}^2 \to \{0,1\}$. O suporte de uma imagem binária Q é um subconjunto finito de $\mathbb{Z}^2$ onde a imagem está de fato definida. O tamanho do suporte é o número de pixels da imagem e uma imagem é considerada estar preenchida com uma cor-de-fundo fora do seu suporte.

Um W-operador binário Ψ é uma função que mapeia uma imagem binária numa outra, definida através de um conjunto de w pontos chamado janela

$$\vec{W} = \{W_1, \ldots, W_w\},\ W_i \in \mathbb{Z}^2$$

e um conceito ou uma função característica

$$\psi: \{0,1\}^w \to \{0,1\}$$

como segue:

$$\Psi(Q)(p) = \psi(Q(W_1 + p), \ldots, Q(W_w + p)),$$

onde $p \in \mathbb{Z}^2$. Cada ponto W_i da janela é chamado *peephole*.

Sejam as imagens $A^{\mathbf{x}}$, $A^{\mathbf{y}}$, $Q^{\mathbf{x}}$ e $Q^{\mathbf{y}}$ respectivamente a imagem de entrada de treinamento, imagem de saída de treinamento, a imagem a ser processada e a imagem de saída ideal (supostamente desconhecida). Podemos supor que existe um único par de imagens de treinamento ($A^{\mathbf{x}}$ e $A^{\mathbf{y}}$), porque se existirem muitos pares, elas podem ser "coladas" para formarem um único par. A fim de projetar W-operator $\hat{\Psi}$, o usuário deve escolher manualmente uma janela apropriada $\vec{W}$.

Vamos denotar o conteúdo em $A^{\mathbf{x}}$, da janela $\vec{W}$ deslocada para $p \in \mathbb{Z}^2$, como $a_p^{\mathbf{x}}$ e denominá-lo uma instância de treinamento ou um padrão de entrada em torno do pixel p:

$$a_p^{\mathbf{x}} = [A^{\mathbf{x}}(W_1 + p), A^{\mathbf{x}}(W_2 + p), \ldots, A^{\mathbf{x}}(W_w + p)] \in \{0,1\}^w.$$

Cada padrão $a_p^{\mathbf{x}}$ está associado com uma cor de saída ou classificação $A^{\mathbf{y}}(p) \in \{0,1\}$. Vamos denotar os dados obtidos quando todos os pixels de $A^{\mathbf{x}}$ e $A^{\mathbf{y}}$ são varridos como uma seqüência

$$\vec{a} = \left((a_{p_1}^{\mathbf{x}}, A^{\mathbf{y}}(p_1)), \ldots, (a_{p_m}^{\mathbf{x}}, A^{\mathbf{y}}(p_m))\right)$$

e denominá-la seqüência de amostras (m é a quantidade de pixels das imagens $A^{\mathbf{x}}$ e $A^{\mathbf{y}}$). Cada elemento $(a_{p_i}^{\mathbf{x}}, A^{\mathbf{y}}(p_i)) \in \vec{a}$ é chamado um exemplo ou uma amostra. Vamos construir de forma semelhante a seqüência

$$\vec{q} = \left((q_{p_1}^{\mathbf{x}}, Q^{\mathbf{y}}(p_1)), \ldots, (q_{p_n}^{\mathbf{x}}, Q^{\mathbf{y}}(p_n))\right)$$

a partir de $Q^{\mathbf{x}}$ e $Q^{\mathbf{y}}$ (n é a quantidade de pixels de $Q^{\mathbf{x}}$ e $Q^{\mathbf{y}}$). Cada $q_{p_i}^{\mathbf{x}}$ é chamado um padrão de busca ou uma instância a ser processada, e a saída $Q^{\mathbf{y}}(p_i) \in \{0,1\}$ é chamada a cor de saída ideal ou a classificação ideal.

O aprendiz ou o algoritmo de aprendizagem **A** é requisitado para construir, baseado em $A^{\mathbf{x}}$ e $A^{\mathbf{y}}$, um W-operator $\hat{\Psi}$ tal que, quando $\hat{\Psi}$ é aplicado à $Q^{\mathbf{x}}$, espera-se que a imagem resultante $\hat{Q}^{\mathbf{y}} = \hat{\Psi}(Q^{\mathbf{x}})$ seja semelhante à imagem de saída ideal $Q^{\mathbf{y}}$. Mais precisamente, o aprendiz **A** deve construir uma função característica ou hipótese $\hat{\psi}$ baseado em seqüência de amostras $\vec{a}$ de forma que, quando $\hat{\psi}$ é aplicado a um padrão de busca $q_{p_i}^{\mathbf{x}}$, espera-se que a sua classificação $\hat{Q}^{\mathbf{y}}(p_i) = \hat{\psi}(q_{p_i}^{\mathbf{x}})$ seja igual a $Q^{\mathbf{y}}(p_i)$ com alta probabilidade. A função $\hat{\psi}$ e a janela $\vec{W}$ juntas representam o W-operator $\hat{\Psi}$.

3. Caso sem Ruído

Vamos estudar em primeiro lugar o caso sem ruído. Pois, embora a maioria dos problemas práticos sejam ruidosos, o estudo do caso sem ruído irá nos ajudar a compreender melhor os casos ruidosos.

Num ambiente sem ruído, existe um conceito alvo claramente definido $\psi: \{0,1\}^w \to \{0,1\}$ que o aprendiz deve aprender. Em tal ambiente, podemos supor que as instâncias de treinamento $a_{p_i}^{\mathbf{x}}$ são geradas aleatória e independentemente no espaço $\{0,1\}^w$ por uma distribuição de probabilidade P. Além disso, as cores de saída $A^{\mathbf{y}}(p_i)$ são geradas aplicando a função alvo ψ em cada $a_{p_i}^{\mathbf{x}}$, isto é, $A^{\mathbf{y}}(p_i) = \psi(a_{p_i}^{\mathbf{x}})$ para todos os pares $(a_{p_i}^{\mathbf{x}}, A^{\mathbf{y}}(p_i)) \in \vec{a}$.

O aprendiz **A** deve considerar algum conjunto $H \subset \left(\{0,1\}^w \to \{0,1\}\right)$ de possíveis hipóteses quando tenta aprender o conceito alvo ψ. Se nenhuma informação sobre ψ estiver disponível, o aprendiz deve assumir que $H = \left(\{0,1\}^w \to \{0,1\}\right)$. Porém, uma informação *a priori* pode simplificar bastante o processo de aprendizagem, pois ela pode reduzir substancialmente a cardinalidade do espaço das hipóteses H. Por exemplo, emular uma erosão Ψ com a informação de que Ψ é uma erosão é muito mais fácil do que emulá-la sem nenhuma informação a priori (exemplos 2 e 3). Uma erosão é um operador elementar de morfologia matemática e a sua definição encontra-se, por exemplo, em [8]. No estágio de treinamento de W-operator, o aprendiz **A** recebe uma seqüência de amostras $\vec{a}$ e procura uma hipótese $\hat{\psi} = \mathbf{A}(\vec{a})$ no espaço H.

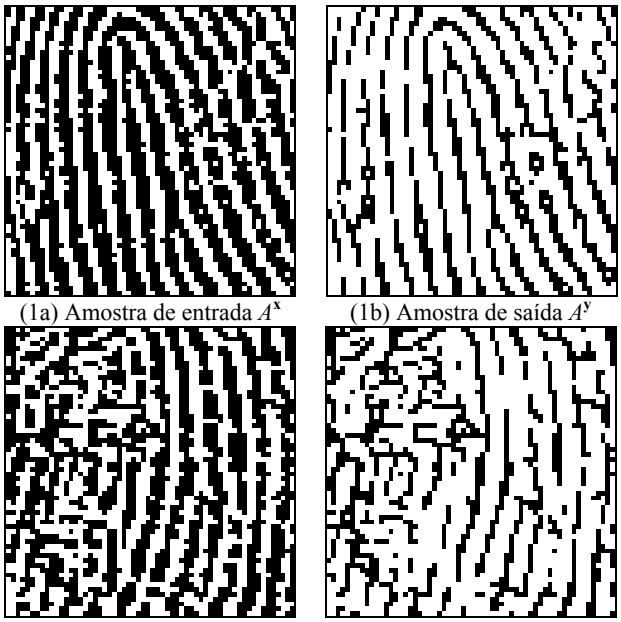
Vamos definir o erro verdadeiro da hipótese $\hat{\psi}$ como a probabilidade de que $\hat{\psi}$ irá classificar incorretamente uma instância $q_{p_i}^{\mathbf{x}}$ escolhida aleatoriamente por P:

$$\text{error}_P(\hat{\psi}) = P\left\{q_{p_i}^{\mathbf{x}} \in \{0,1\}^w \;\middle|\; \psi(q_{p_i}^{\mathbf{x}}) \neq \hat{\psi}(q_{p_i}^{\mathbf{x}})\right\}$$

De acordo com a teoria PAC [9, 10], qualquer aprendiz consistente utilizando um espaço de hipótese finito H com função alvo $\psi \in H$ irá, com probabilidade maior que (1-δ), gerar uma hipótese $\hat{\psi}$ com erro menor que ε, depois de observar m exemplos escolhidos aleatoriamente pelo P, desde que

$$m \geq \frac{1}{\varepsilon}\left[\ln\left(\frac{1}{\delta}\right) + \ln\left(|H|\right)\right]. \qquad (1)$$

Um aprendiz é consistente se, sempre que possível, gerar uma hipótese que se adapte perfeitamente aos dados de treinamento. O limite (1) freqüentemente está substancialmente superestimado, principalmente porque nenhuma suposição foi feita sobre o aprendiz exceto a consistência. Alguns exemplos de uso desta equação seguem.



(1a) Amostra de entrada $A^{\mathbf{x}}$ (1b) Amostra de saída $A^{\mathbf{y}}$

(1c) A ser processado $Q^{\mathbf{x}}$ (1d) Saída ideal $Q^{\mathbf{y}} = \hat{Q}^{\mathbf{y}}$

Figura 1: Aprendizagem de W-operator em ambiente sem ruído.

Exemplo 1: Na figura 1, uma imagem de impressão digital $A^{\mathbf{x}}$ (1a) foi processada por W-operator Ψ, gerando a imagem $A^{\mathbf{y}}$ (1b). Este operador consistiu em união de 8 operadores hit-or-miss definidos dentro da janela 3×3. O operador hit-or-miss é um dos operadores elementares da morfologia matemática e a sua definição encontra-se, por exemplo, em [8]. Vamos supor que de alguma forma conhecemos que Ψ está definida na janela 3×3. Utilizando esta informação e as imagens $A^{\mathbf{x}}$ e $A^{\mathbf{y}}$, um W-operator $\hat{\Psi}$ foi construído por um aprendiz consistente. De acordo com a equação (1), com probabilidade maior que 99%, o erro verdadeiro de $\hat{\Psi}$ será menor que 1%, desde que as imagens de treinamento tenham uma quantidade de pixels

$$m \geq \frac{1}{0.01}\left[\ln\left(\frac{1}{0.01}\right) + \ln\left(2^{2^9}\right)\right] \cong 35950 .$$

Como as imagens $A^{\mathbf{x}}$ e $A^{\mathbf{y}}$ têm 200×200=40000 pixels (somente pequenas partes das imagens originais são mostradas na figura 1), quase certamente $\hat{\Psi}$ irá apresentar uma taxa de erro menor que 1%. Sem surpresas, quando $\hat{\Psi}$ foi aplicado a uma outra imagem de impressão digital (figura 1c), uma image $\hat{Q}^{\mathbf{y}} = \hat{\Psi}(Q^{\mathbf{x}})$ (figura 1d) exatamente igual à saída ideal $Q^{\mathbf{y}} = \Psi(Q^{\mathbf{x}})$ foi produzida, isto é, $\hat{\Psi}$ apresentou erro zero. Este teste foi repetido algumas vezes e as taxas de erro sempre foram zero. ■

Note que a análise acima somente é válida quando se pode supor que as imagens $A^{\mathbf{x}}$ e $Q^{\mathbf{x}}$ foram geradas por uma mesma distribuição de probabilidade. Isto é, $A^{\mathbf{x}}$ e $Q^{\mathbf{x}}$ devem ser de um mesmo tipo: imagens de impressões digitais, documentos manuscritos, documentos impressos, etc.

Exemplo 2: Vamos resolver novamente o exemplo 1, desta vez supondo que o operador alvo é mais complexo e está definido dentro de uma janela 7×7. Neste caso:

$$m \geq \frac{1}{0.01}\left[\ln\left(\frac{1}{0.01}\right) + \ln\left(2^{2^{49}}\right)\right] \cong 3.9 \times 10^{16} .$$

Isto é, as imagens amostras devem ser maiores que $(2\times10^8)\times(2\times10^8)$! Claramente, uma imagem tão grande não pode ser obtida na prática. ■

Exemplo 3: Vamos resolver novamente o exemplo 2, desta vez supondo que temos conhecimento de que o operador alvo é uma erosão cujo elemento estruturante cabe dentro de uma janela 7×7. Como cada um dos 49 *peepholes* pode pertencer ou não ao elemento estruturante, o operador alvo tem de ser uma das 2^{49} erosões. Assim, $|H| = 2^{49}$ e:

$$m \geq \frac{1}{0.01}\left[\ln\left(\frac{1}{0.01}\right) + \ln\left(2^{49}\right)\right] \cong 3857 .$$

Isto é, qualquer par de imagens de treinamento maiores que 63×63 será suficiente. Compare com o tamanho das imagens $(2\times10^8)\times(2\times10^8)$ do exemplo 2. ■

A simplificação acima somente é válida quando se utiliza um algoritmo de aprendizagem projetado especialmente para erosões. Resultados semelhantes podem ser obtidos para outros operadores elementares tais como dilatação, hit-or-miss, união de *k* erosões, e assim por diante [9, 10].

4. Caso Ruidoso

Para modelar o caso ruidoso, vamos supor que cada exemplo $(a_p^{\mathbf{x}}, A^{\mathbf{y}}(p)) \in \vec{a}$ tenha sido gerado independentemente por uma distribuição de probabilidade conjunta *P* desconhecida no espaço $\{0,1\}^w \times \{0,1\}$. Vamos também supor que cada elemento $(q_{p_i}^{\mathbf{x}}, Q^{\mathbf{y}}(p_i)) \in \vec{q}$ tenha sido gerado pela mesma distribuição *P*.

O erro verdadeiro da hipótese ψ agora deve ser definido como a probabilidade de que ψ classifique incorretamente um exemplo $(q_{p_i}^{\mathbf{x}}, Q^{\mathbf{y}}(p_i))$ escolhido aleatoriamente por *P*:

$$\text{error}_P(\psi) = P\left\{(q_{p_i}^{\mathbf{x}}, Q^{\mathbf{y}}(p_i)) \in \{0,1\}^w \times \{0,1\} \;\middle|\; \psi(q_{p_i}^{\mathbf{x}}) \neq Q^{\mathbf{y}}(p_i)\right\}$$

Na situação ruidosa, não existe uma função alvo claramente definida. No seu lugar, existe uma função ψ^* com o menor erro verdadeiro. Vamos definir o erro empírico (e-erro) de uma hipótese ψ sobre uma seqüência $\vec{a}$ como a proporção de erros cometidos quando ψ classifica as instâncias de $\vec{a}$:

$$\text{error}_{\vec{a}}(\psi) = \frac{1}{m}\left|\left\{(a_{p_i}^{\mathbf{x}}, A^{\mathbf{y}}(p_i)) \in \vec{a} \;\middle|\; \psi(a_{p_i}^{\mathbf{x}}) \neq A^{\mathbf{y}}(p_i)\right\}\right| ,$$

onde *m* é o comprimento de $\vec{a}$.

Seja $\hat{\psi}$ a hipótese com o menor e-erro sobre $\vec{a}$ e seja ψ^* a hipótese com o menor erro verdadeiro. Então [11]

$$\Pr\left[\text{error}_P(\hat{\psi}) - \text{error}_P(\psi^*) > \varepsilon\right] < \delta ,$$

desde que *H* seja finito e o comprimento *m* de $\vec{a}$ satisfaça:

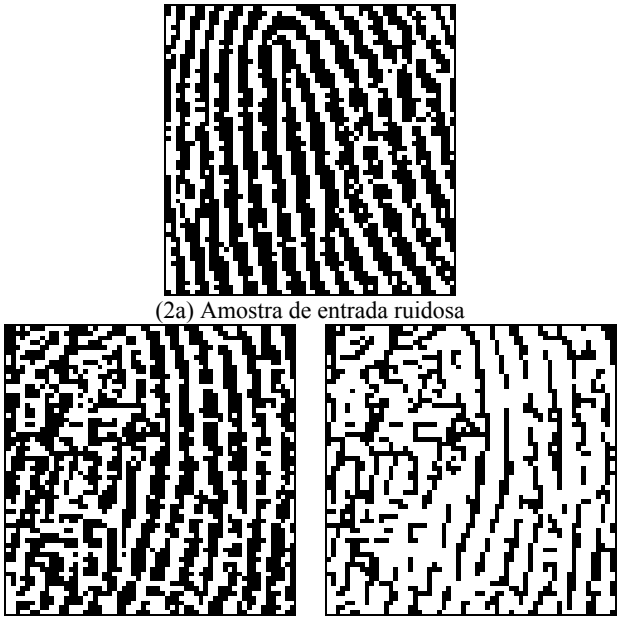
$$m \geq \frac{1}{2\varepsilon^2}\left[\ln\left(\frac{1}{\delta}\right) + \ln\left(2|H|\right)\right]. \qquad (2)$$

Infelizmente, a complexidade de amostra acima é uma superestimativa ainda maior que a da equação (1). Dada uma seqüência de amostras $\vec{a}$, a hipótese empiricamente ótima (e-ótima) $\hat{\psi}$ pode ser construída facilmente. Vamos definir que um aprendiz **A** é e-ótimo se ele gerar sempre uma hipótese e-ótima sobre a seqüência de treinamento. Se **A** fosse e-ótimo, dado um padrão de busca $q_{p_i}^{\mathbf{x}}$, qual deveria ser a sua classificação $\hat{\psi}(q_{p_i}^{\mathbf{x}}) = \mathbf{A}(\vec{a})(q_{p_i}^{\mathbf{x}})$? Sejam $(a_{r_1}^{\mathbf{x}}, A^{\mathbf{y}}(r_1))$, ...,

$(a^{\mathbf{x}}_{r_N}, A^{\mathbf{y}}(r_N))$ os N exemplos de treinamento de $q^{\mathbf{x}}_{p_i}$ em $\vec{a}$, isto é, $a^{\mathbf{x}}_{r_j} = q^{\mathbf{x}}_{p_i}$, $1 \le j \le N$ (não há outros exemplos de $q^{\mathbf{x}}_{p_i}$ em $\vec{a}$ além desses). Como há ruído, os N exemplos acima podem não concordar sobre a classificação de $q^{\mathbf{x}}_{p_i}$. Para minimizar e-erro, a classificação deve ser decidida pela maioria dos votos desses exemplos de treinamento:

$$\hat{\psi}(q^{\mathbf{x}}_{p_i}) \leftarrow \text{mode}(A^{\mathbf{y}}(r_1),\ldots,A^{\mathbf{y}}(r_N)) .$$

Note que todo aprendiz e-ótimo é consistente num ambiente sem ruído. Apresentamos abaixo um exemplo.



(2a) Amostra de entrada ruidosa

(2b) Imagem ruidosa a processar

(2c) Imagem processada

Figura 2: Aprendizagem de W-operator em ambiente ruidoso.

Exemplo 4: As imagens de impressões digitais 1a e 1c foram corrompidas pelo ruído "sal e pimenta", resultando em imagens 2a e 2b. Em média, 1 em cada 40 pixels mudaram de cor. Gostaríamos de projetar um W-operador 3×3 $\hat{\Psi}$ tal que uma imagem semelhante à saída ideal $A^{\mathbf{y}}$ (figura 1b) resulte, apesar do ruído, quando a imagem 2a é processada por $\hat{\Psi}$. Para atingir este objetivo, um W-operator $\hat{\Psi}$ foi projetado por um aprendiz e-ótimo usando imagens 2a e 1b como amostras de treinamento. Como as imagens 2a e 1b têm 200×200 pixels, com probabilidade pelo menos 99%, a diferença entre os erros verdadeiros do operador 3×3 ótimo Ψ^* e do $\hat{\Psi}$ será menor que 6.71%, i.e., $\text{error}_P(\hat{\psi}) - \text{error}_P(\psi^*) \le 0.0671$, pois:

$$\frac{1}{2\times 0.0671^2}\left[\ln\left(\frac{1}{0.01}\right)+\ln\left(2\times 2^{2^9}\right)\right] \cong 40000 .$$

No exemplo 5, este problema será analisado novamente. ■

5. Estimação Estatística da Taxa de Erro

Esta subseção irá expor as técnicas para calcular um limite mais estreito para a taxa de erro. Estas técnicas serão muito úteis, pois as equações (1) e (2) normalmente superestimam a complexidade de amostra e a taxa de erro. Ao contrário das fórmulas anteriores, as técnicas desta subseção podem ser aplicadas somente após ter projetado W-operador, com a condição adicional de que a imagem de saída ideal $Q^{\mathbf{y}}$ esteja disponível. É lícito supor que a saída ideal estará disponível para se realizar testes, pois estamos supondo que um par de imagens entrada-saída de treinamento está disponível para projetar W-operador. E se as imagens de treinamento estão disponíveis, elas podem ser quebradas em dois pedaços: imagens de treinamento ($A^{\mathbf{x}}$, $A^{\mathbf{y}}$) e imagens de teste ($Q^{\mathbf{x}}$, $Q^{\mathbf{y}}$).

Portanto, supondo que a saída ideal $Q^{\mathbf{y}}$ esteja disponível, uma simples contagem de pixels diferentes entre $Q^{\mathbf{y}}$ e $\hat{Q}^{\mathbf{y}}$ irá fornecer o e-erro. E, dada a acuracidade observada de uma hipótese sobre uma amostra de dados limitada, é possível conhecer o quanto esta irá conseguir estimar a acuracidade sobre exemplos adicionais. Para isso, vamos construir intervalos de confiança unilateral ou bilateral. Explicações adicionais sobre intervalos de confiança da média de variáveis aleatórias binomiais encontram-se em [9] ou em muitos livros elementares de Estatística. Com N% de confiança:

$$\text{error}_P(\hat{\psi}) \in \text{error}_{\vec{q}}(\hat{\psi}) \pm z_N\sqrt{\frac{\text{error}_{\vec{q}}(\hat{\psi})(1-\text{error}_{\vec{q}}(\hat{\psi}))}{n}} , \quad (3)$$

$$\text{error}_P(\hat{\psi}) \le \text{error}_{\vec{q}}(\hat{\psi}) + z'_N\sqrt{\frac{\text{error}_{\vec{q}}(\hat{\psi})(1-\text{error}_{\vec{q}}(\hat{\psi}))}{n}} , \quad (4)$$

onde n é o comprimento de $\vec{q}$; z_N define a metade da largura do menor intervalo em torno da média que inclui N% da massa da probabilidade total sob distribuição normal com desvio-padrão 1; e $z'_N \equiv z_{2N-1}$. Por exemplo, $z'_{84\%} = z_{68\%} = 1.00$, $z'_{95\%} = z_{90\%} = 1.64$, $z'_{99\%} = z_{98\%} = 2.33$ e $z_{99\%} = 2.58$. As fórmulas (3) e (4) normalmente produzem uma estimativa da taxa de erro muito mais acurada que as equações (1) e (2).

No caso sem ruído, basta conhecer um limite superior para a taxa de erro verdadeiro do operador projetado ($\text{error}_P(\hat{\psi})$). Porém, para casos ruidosos, o erro mínimo ($\text{error}_P(\psi^*)$) também deve ser estimado pois, como o operador projetado nunca poderá atingir uma taxa de erro verdadeiro menor que o mínimo, um operador pode ser considerado uma boa solução se o seu erro verdadeiro estiver próximo do mínimo. Infelizmente, não há meios para se estimar $\text{error}_P(\psi^*)$ diretamente, pois o otimador ótimo é desconhecido. Descrevemos abaixo um artifício que tem conseguido estabelecer bons limites inferiores para $\text{error}_P(\psi^*)$. Embora muito simples, nunca vimos esta técnica descrita na literatura.

Para isso, vamos construir a hipótese $\hat{\psi}^*$ e-ótima sobre $\vec{q}$. Se o aprendiz **A** for e-ótimo, $\hat{\psi}^* = \mathbf{A}(\vec{q})$. Note que estamos treinando o operador com as próprias imagens ($Q^{\mathbf{x}}$, $Q^{\mathbf{y}}$) que serão utilizadas no teste. Claramente, $\text{error}_{\vec{q}}(\hat{\psi}^*) \le \text{error}_{\vec{q}}(\psi^*)$ e $\text{error}_{\vec{q}}(\hat{\psi}^*)$ pode ser medido experimentalmente. Então, utilizamos a seguinte desigualdade para estabelecer um limite inferior para $\text{error}_P(\psi^*)$:

$$\begin{aligned}\text{error}_P(\psi^*) &\\ &\ge \text{error}_{\vec{q}}(\psi^*) - z'_N\sqrt{\frac{\text{error}_{\vec{q}}(\psi^*)(1-\text{error}_{\vec{q}}(\psi^*))}{n}} \quad (5)\\ &\ge \text{error}_{\vec{q}}(\hat{\psi}^*) - z'_N\sqrt{\frac{\text{error}_{\vec{q}}(\hat{\psi}^*)(1-\text{error}_{\vec{q}}(\hat{\psi}^*))}{n}}\end{aligned}$$

A desigualdade acima é verdadeira, com nível de confiança N%, toda vez que:

$$\frac{b+1-\sqrt{b(b+1)}}{2(b+1)} \le \text{error}_{\vec{q}}(\hat{\psi}^*) \le \text{error}_{\vec{q}}(\psi^*) \le \frac{b+1+\sqrt{b(b+1)}}{2(b+1)} \quad (6)$$

onde $b = n/(z'_N)^2$. Note que a desigualdade (6) é verdadeira para praticamente todos problemas práticos e conseqüentemente a desigualdade (5) também é sempre verdadeira na prática.

Exemplo 5: No exemplo 4, concluímos com 99% de confiança que o operador $\hat{\Psi}$ obtido comete no máximo 6.71% mais erros que o operador ótimo 3×3. A fim de estabelecer um limite de erro mais estreito, e-erro de $\hat{\Psi}$ (a diferença entre as imagens 1d e 2c) foi medido e descobriu-se que valia 4.992%. Utilizando a equação (3), concluímos com 99% de confiança que o erro verdadeiro de $\hat{\Psi}$ pertence ao intervalo (4.992±0.281)%. O operador $\hat{\Psi}^*$ e-ótimo sobre imagens testes (figuras 2b e 1d) foi construído e cometeu e-erro 4.723% quando processou a imagem 2b. Utilizando a desigualdade (5), concluímos com 99% de confiança que o erro verdadeiro do operador ótimo 3×3 é maior que (4.723-0.247)%. Conseqüentemente, com confiança de pelo menos 99%, o erro verdadeiro de $\hat{\Psi}$ é no máximo 0.797% maior que o erro verdadeiro do operador ótimo, isto é:

$$\text{error}_P(\hat{\psi}) - \text{error}_P(\psi^*) \le 0.00797 .$$

Este resultado confirma que a equação (2) superestima a taxa de erro, pois 0.797% é muito menor que 6.71%. ■

6. Generalização

Nas seções 3 e 4, supusemos que o aprendiz é e-ótimo (ou consistente) para calcular a complexidade de amostra. Porém a e-otimalidade não especifica inteiramente um algoritmo de aprendizagem, pois existem muitos diferentes aprendizes e-ótimos. Para especificar completamente um aprendiz, um método de generalização (viés indutivo) também deve ser escolhido.

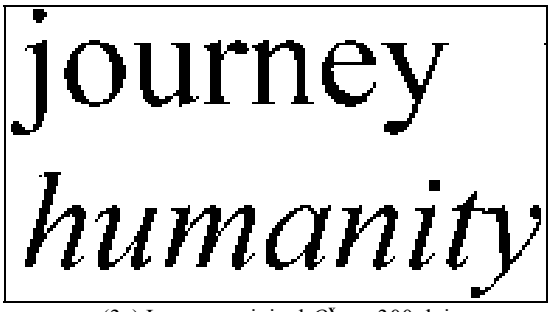
Para a aprendizagem de W-operador, sugerimos que se utilize a generalização *k* vizinhos mais próximos [9], pois parecenos bastante natural que padrões semelhantes sejam classificados similarmente. Outra possibilidade seria utilizar a generalização dada pela árvore de decisão [9], pois se aproxima muito da generalização *k* vizinhos mais próximos. Evidentemente, ao se escolher um viés indutivo, deve-se levar em conta a existência de algoritmos computacionalmente eficientes que consigam implementá-lo. Também deve-se tomar cuidado para que a generalização mantenha a e-otimalidade, pois caso contrário a teoria PAC pode se tornar inválida.

As eficácias dos diferentes vieses indutivos podem ser comparados utilizando a estimação estatística. Sugerimos o livro [9] para maiores detalhes.

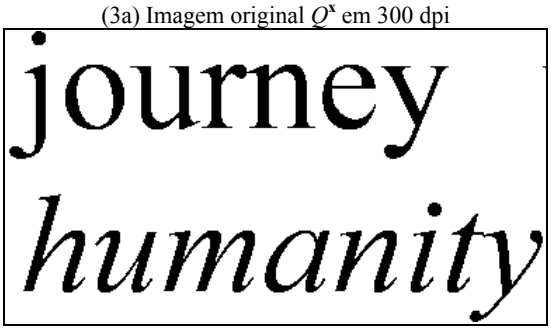
7. Uma Aplicação

Nesta seção iremos aplicar a teoria desenvolvida até agora para analisar a complexidade de amostra e a taxa de erro do problema de aumentar a resolução de documentos impressos por um fator *f* inteiro. Por exemplo, fator *f*=2 deverá aumentar a resolução duas vezes tanto em linhas quanto em colunas. Este problema é inteiramente análogo ao problema de projetar W-operador, exceto que f^2 funções características devem ser projetadas pela aprendizagem. Remetemos o leitor para [7] para maiores detalhes da descrição deste problema e a sua solução computacional. Aqui estamos interessados somente no seu aspecto estatístico. Um operador que aumenta a resolução será denotado como WZ-operador (Z de *zoom*).

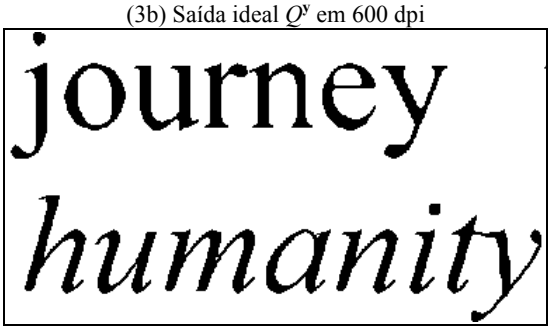
Projetamos um WZ-operador 3×3 para aumentar a resolução de documentos contendo caracteres "Times 12 pt." (tanto normal como itálico) de 300 dpi para 600 dpi. As imagens de treinamento foram obtidas imprimindo documentos eletrônicos para arquivos através de um *driver* de impressora *PostScript*, e então convertendo esses arquivos para imagens binárias. Embora as imagens estejam livres de ruído, o problema deve ser considerado ruidoso, pois um único padrão 3×3 em 300 dpi pode corresponder a dois ou mais padrões diferentes em 600 dpi.



(3a) Imagem original Q^x em 300 dpi



(3b) Saída ideal Q^y em 600 dpi



(3c) Imagem 600 dpi $\hat{Q}^y$ gerada pela aprendizagem

Figura 3: Aumento de resolução de caracteres impressos.

Vamos utilizar a equação 2 para estimar o tamanho das imagens de treinamento necessárias para, usando janela 3×3, obter WZ-operator $\hat{\Psi}$ com uma taxa de erro no máximo 2% maior que o operador ótimo. Usando nível de confiança 99%:

$$m \ge \frac{1}{2\varepsilon^2}\left[\ln\left(\frac{1}{\delta}\right) + \ln\left(2|H|\right)\right] = \frac{1}{2\times 0.02^2}\left[\ln\left(\frac{1}{0.01}\right) + \ln(2) + 2^9 \times \ln(2)\right] \cong 450238$$

Temos dois pares de imagens de amostra independentes ($A^{\mathbf{x}}$, $A^{\mathbf{y}}$) e ($Q^{\mathbf{x}}$, $Q^{\mathbf{y}}$) com caracteres Times 12 pt. (figura 3) cujos tamanhos são (554×813, 1108×1626) e (558×740, 1116×1480), respectivamente. Note que a imagem $A^{\mathbf{x}}$ é grande o suficiente para obter a acuracidade desejada, pois 554×813=450402. Um WZ-operador foi construído utilizando a aprendizagem 1-vizinho mais próximo.

A imagem processada $\hat{Q}^{\mathbf{y}}$ (figura 3c) e a imagem ideal $Q^{\mathbf{y}}$ (figura 3b) diferiam em 1.058% dos pixels. Uma vez que o e-erro foi medido, a seguinte pergunta pode surgir: "É possível aumentar substancialmente a acuracidade do operador projetado?" Utilizaremos as desigualdades (4) e (5) para mostrar que é impossível obter qualquer melhora substancial na qualidade do WZ-operador, enquanto janela 3×3 estiver sendo utilizada. Nós mostraremos que:

1) O e-erro obtido é uma boa estimativa do erro verdadeiro de $\hat{\Psi}$.

2) O erro verdadeiro do operador 3×3 ótimo está muito próximo ao do $\hat{\Psi}$.

Usando a equação (4), com confiança 99%:

$$\begin{aligned}\text{error}_P(\hat{\psi}) &\le 0.01058 + 2.33\sqrt{\frac{0.01058(1-0.01058)}{1116\times 1480}} \\ &= (1.058 + 0.019)\%\end{aligned}$$

o que demonstra a primeira afirmação.

Para demonstrar a segunda afirmação, criamos o WZ-operador $\hat{\Psi}^*$ e-ótima sobre ($Q^{\mathbf{x}}$, $Q^{\mathbf{y}}$) e o aplicamos à imagem $Q^{\mathbf{x}}$. The e-erro obtido foi 1.045%. Utilizando os dados obtidos e a equação (5), concluímos com 99% de confiança que:

$$\begin{aligned}\text{error}_P(\psi^*) &\ge 0.01045 - 2.33\sqrt{\frac{0.01045(1-0.01045)}{1116\times 1480}} \\ &= (1.045 - 0.018)\%\end{aligned}.$$

Isto mostra claramente que não pode existir qualquer WZ-operador 3×3 substancialmente melhor que $\hat{\Psi}$, pois com probabilidade 99% o erro verdadeiro de $\hat{\Psi}$ é no máximo 1.077% enquanto que com a mesma probabilidade o erro verdadeiro do operador 3×3 ótimo é no mínimo 1.027%.

Uma vez que nós demonstramos que o WZ-operador obtido é virtualmente o melhor WZ-operador 3×3, uma outra questão pode surgir: "Poderia melhorar a qualidade do operador escolhendo uma janela maior"? Repetimos os testes utilizando uma janela com 17 *peepholes* e obtivemos e-erro 0.995%. Isto mostra que a qualidade pode melhorar aumentando a janela, mas aparentemente de forma lenta com o aumento da janela. Na realidade, necessitamos de mais testes para responder a esta pergunta de uma forma definitiva.

8. Conclusão

Neste trabalho, descrevemos como a teoria de aprendizagem PAC (Provavelmente Aproximadamente Correta) pode ser utilizada para estimar a complexidade de amostra do problema de aprendizagem de operadores para imagens binárias. Descrevemos a técnica tanto para o caso sem ruído quanto para o caso ruidoso. Como esta teoria costuma fornecer uma complexidade de amostra superestimada, descrevemos o uso de estimação estatística para calcular, *a posteriori*, uma taxa de erro estreita. Também mostramos como se pode estimar a taxa de erro mínima de um problema ruidoso. Aplicamos a teoria desenvolvida para analisar o problema de aumento de resolução espacial de caracteres impressos.

Referências

[1] E. R. Dougherty, "Optimal Mean-Square N-Observation Digital Morphological Filters, Part II - Optimal Gray-Scale Filters," *CVGIP: Image Understanding*, vol. 55, no. 1, pp. 55-72, 1992.

[2] H. Y. Kim, "Quick Construction of Efficient Morphological Operators by Computational Learning," *Electronics Letters*, vol. 33, no. 4, pp. 286-287, 1997.

[3] H. Y. Kim, and F. A. M. Cipparrone, "Automatic Design of Nonlinear Filters by Nearest Neighbor Learning," In *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, pp. 737-741, TP7.05, 1998.

[4] H. Y. Kim, "Segmentation-Free Printed Character Recognition by Relaxed Nearest Neighbor Learning of Windowed Operator," In *Proc. Brazilian Symp. on Comp. Graph. and Image Proc.*, pp. 195-204, 1999.

[5] E. R. Dougherty, "Optimal Mean-Square N-Observation Digital Morphological Filters, Part I - Optimal Binary Filters," *CVGIP: Image Understanding*, vol. 55, no. 1, pp. 36-54, 1992.

[6] R. P. Loce, E. R. Dougherty, R. E. Jodoin, and M. S. Cianciosi, "Logically Efficient Spatial Resolution Conversion Using Paired Increasing Operators," *Real-Time Imaging*, vol. 3, no. 1, pp. 7-16, 1997.

[7] H. Y. Kim, and P. S. L. M. Barreto, "Fast Binary Image Resolution Increasing by *k*-Nearest Neighbor Learning," In *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, pp. 327-330, TA9.06, 2000.

[8] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, 1992.

[9] T. M. Mitchell, *Machine Learning*, WCB/McGraw-Hill, 1997.

[10] M. Anthony and N. Biggs, *Computational Learning Theory - An Introduction*, Cambridge University Press, 1992.

[11] D. Haussler, "Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications," *Information and Computation*, vol. 100, 1992, pp. 78-150.