Estimação dos Parâmetros de uma SVM utilizando um Algoritmo Genético para o Reconhecimento de Caracteres Manuscritos

F. Wilson R. Júnior[1], **Kennedy S. de Abreu**[1]

[1]Universidade Federal do Ceará (UFC)
Sobral – Ce – Brasil

wilsoonjunior@gmail.com, kennedysouza11994@hotmail.com

Abstract. *This work deals with an HCR system using a genetic algorithm (GA) to estimate the kernel parameters of an SVM for the recognition of the MNIST database in order to increase the correctness percentage of the classifier with the best combination of parameters . Also discussed are digital image processing (PDI) techniques for image processing and feature extraction. The use of a GA together with an SVM proved to be effective, achieving high accuracy rates.*

Resumo. *Este trabalho aborda um sistema HCR utilizando um algoritmo genético (AG) para estimar os parâmetros do kernel de uma SVM para o reconhecimento dos caracteres manuscritos da base de dados MNIST com o objetivo de elevar o percentual de acertos do classificador com a melhor combinação dos parâmetros. Também são abordadas técnicas de processamento digital de imagens (PDI) para o tratamento das imagens e extração de características. A utilização de um AG em conjunto com uma SVM mostrou-se eficaz, alcançando elevadas taxas de precisão.*

1. Introdução

O reconhecimento de caracteres manuscritos ou *handwritten character recognition* (HCR) é uma importante e desafiadora área de estudos de PDI e reconhecimento de padrões, utilizado para diversos fins como, reconhecimento de placas de veículos, preservação de documentos históricos manuscritos e reconhecimento de endereços [Pradeep et al. 2012],[Mori et al. 1995].

O HCR pode ser realizado por meio de técnicas de inteligência computacional de treinamento supervisionado, ou não supervisionado. O treinamento supervisionado é capaz de aprender a classificar conjuntos de padrões rotulados com suas respectivas classes e partir destes, ser capaz de identificar padrões desconhecidos, como é o caso das técnicas de Rede Neural Artificial (RNA), Máquinas de Vetores de Suporte ou *Support Vector Machine* (SVM) ou Regra dos Vizinhos mais Próximos ou *K-Nearest Neighboors* (KNN).

Neste trabalho é abordado um classificador SVM, devido aos bons resultados encontrados em trabalhos na literatura, como em [Kumar et al. 2012],[Yerra et al. 2017],[Bonesso 2013]. Mas a determinação dos seus parâmetros não é trivial, podendo demandar um tempo elevado na obtenção dos parâmetros ideais, que maximizem o percentual de acertos. Dessa forma, uma solução

viável para esse problema é a utilização de um algoritmo genético real, por ser bastante utilizado para buscas e otimização de problemas complexos a partir de uma função objetivo baseado na teoria da evolução de Charles Darwin [Linden 2012].

Então, motivado pelas inúmeras aplicações em que o HCR pode se inserir, o foco deste trabalho está na construção de um sistema HCR, utilizando técnicas de PDI e de inteligência computacional. O objetivo principal é avaliar o desempenho do sistema HCR com a utilização de um algoritmo genético integrado ao processo de classificação, para estimação automática dos parâmetros de uma SVM com função de kernel *Radial Basis Function* (RBF).

Este trabalho está organizado em seções. Na seção 2, é apresentado uma revisão bibliográfica sobre o tema deste trabalho. Na seção 3 são descritos os principais conceitos e técnicas abordadas. A seção 4 apresenta os principais resultados alcançados e em seguida, na seção 5, é apresentado uma visão geral dos resultados obtidos com a metodologia proposta.

2. Trabalhos Relacionados

Em [Bonesso 2013], é realizado o processo de estimação dos parâmetros do kernel RBF de uma SVM, C e Gamma (γ), utilizando a heurística Recozimento Simulado e Busca em Grade, ambas as técnicas adaptadas para o classificador SVM estruturado com uma árvore binária. A primeira apresentou melhores resultados percorrendo um vasto espaço de busca sem demandar muito tempo.

Em [Rodrigues et al. 2001], são utilizadas técnicas de projeção de contornos baseadas em polígonos regulares para extração de características de dígitos manuscritos. Uma RNA é utilizada para o processo de classificação dos dígitos manuscritos, alcançando uma acurácia máxima de 94,64%.

Em [Coelho 2013], utiliza-se um algoritmo genético binário para otimizar o processo de classificação. Essa otimização ocorre com o algoritmo genético selecionando as melhores características de modo a reduzir o erro de classificação.

3. Metodologia

Nesta seção são apresentados os principais conceitos e processos utilizados para a estimação dos parâmetros do kernel de uma SVM com um algoritmo genético e para o HCR. A Figura 1 apresenta o fluxograma do sistema HCR proposto. Inicialmente, obtém-se as imagens da base de dados *MNIST*, para em seguida aplicar processamento dessas imagens. Posteriormente, determina-se o vetor de características com amostras de treinamento e de teste, sendo apresentado ao algoritmo genético, para determinar a melhor combinação dos parâmetros do kernel da SVM, para assim maximizar a precisão do HCR.

3.1. Aquisição de Dados

A base de dados *MNIST* possui 70000 mil caracteres manuscritos isolados e centralizados em imagens de tamanho 28 por 28 pixels. Encontra-se disponível gratuitamente na web [LeCun et al. 1998] e apresenta grande diversidade nos estilos e formas de escrita dos caracteres. A Figura 2 apresenta algumas amostras da base de dados *MNIST*.

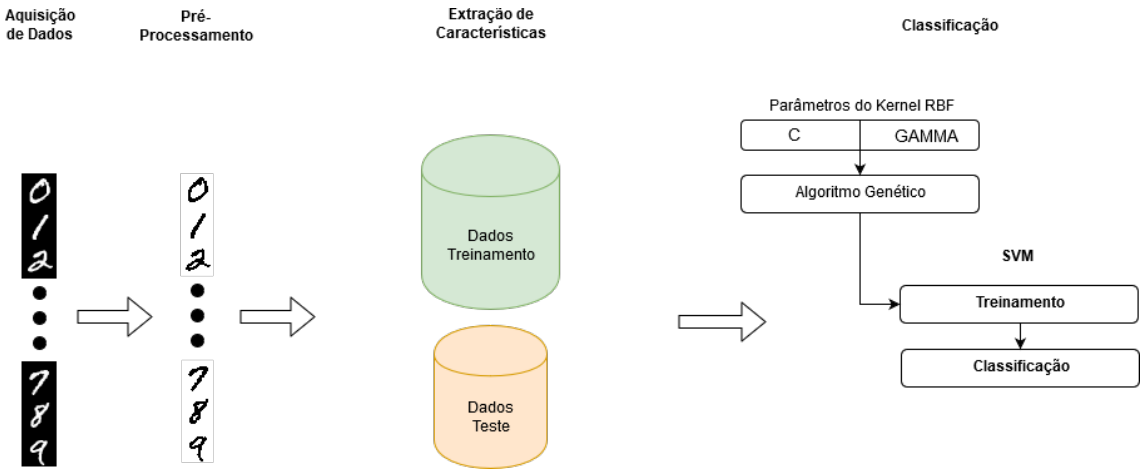

Figura 1. Principais processos do sistema HCR proposto.

Figura 2. Amostras retiradas da base *MNIST*.

3.2. Pré-Processamento

As imagens da base de dados *MNIST* encontram-se com diversos tons de cinza, onde os pixels correspondentes aos caracteres estão com tonalidades mais próximas de branco. Para reduzir o espaço de cores possíveis das imagens, foram utilizadas as técnicas de binarização de imagens, aplicando um *threshold* de 150 e transformação de negativo, ambas descritas em [Gonzalez and Woods 2006].

3.3. Extração de Características

Diversos estudos relacionados a utilização de técnicas de extração de características podem ser encontrados na literatura para obtenção de padrões dos objetos das imagens, direcionados a redução do custo computacional e aumento da eficiência do processo de classificação.

Neste trabalho, foram utilizadas as técnicas de projeção dos contornos, contribuindo com 112 características, e dos histogramas, contribuindo com 56 características, ambas descritas em [Rodrigues et al. 2001] e [Trier et al. 1996] respectivamente.

3.4. Máquina de Vetores de Suporte (SVM)

O conceito de uma SVM é conseguir transformar conjuntos de dados não linearmente separáveis em linearmente separáveis, que por meio de uma função de kernel eleva a dimensão desses conjuntos de dados para serem trabalhados. As SVM's deve ser configurada com uma função de kernel para realizar esse mapeamento dimensional, sendo as mais utilizadas as funções Polinomial, Linear e *Radial Basis Function* (RBF). A última, tem apresentado resultados mais significativos em relação ao problema de HCR [Yerra et al. 2017], [Kumar et al. 2012], [Bonesso 2013] e devido isso foi abordada nesse trabalho.

Outro aspecto relevante para configuração é a estratgia de classificação utilizada em problemas relacionados a muitas classes, pelo fato das SVM's serem classificadores binários. A estratégia de classificação abordada foi a *One Against One* ou Um Contra Um, que construirá um classificador para cada par de classes durante o processo de treinamento dos dados e para classificar dados na etapa de testes, é realizado uma votação onde a classe mais votada é selecionada.

Assim, uma SVM com função de kernel RBF possui dois parâmetros fundamentais para o processo de classificação das características, o C e o γ. O primeiro é responsável por determinar o melhor ponto para separação entre duas classes, enquanto que o segundo determina a largura do kernel ou a influência dos vetores de suporte. A combinação desses parâmetros influencia diretamente o resultado da classificação, porém os encontrar os valores ideais pode ser complexo ou demandar muito tempo no ajuste.

3.5. Algoritmo Genético

A fim de otimizar e melhorar o resultado da classificação das amostras e determinar automaticamente a melhor combinação possível dos parâmetros C e γ, é utilizado um algoritmo genético real por esses parâmetros serem contínuos , assim como em [Bonesso 2013], porém adaptado para o HCR. A Figura 3 representa o fluxograma de um algoritmo genético utilizando elitismo.

Inicialmente define-se o espaço de busca inicial e formato dos cromossomos a serem trabalhados. Em seguida são selecionados aleatoriamente indivíduos da população, utilizando métodos como seleção por roleta ou torneio por exemplo [Linden 2012], baseados em uma função de avaliação, que determina o quão apto a solução é para o problema. Então, esses indivíduos iniciam a reprodução, gerando novos filhos. Por fim, esses filhos passam por processos de mutação com o intuito de conseguir manter a diversidade nessa população e assim os melhores filhos são integrados a população inicial, caso as condições não sejam satisfeitas ao final do elitismo.

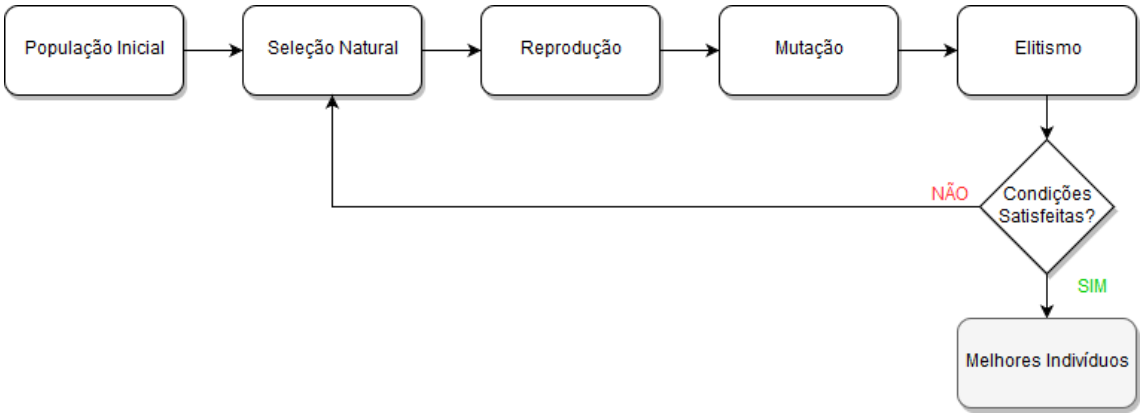

Figura 3. Fluxograma do algoritmo genético.

4. Resultados

Foram utilizadas os 70000 caracteres da base de dados *MNIST* em três experimentos, onde inicialmente foram utilizadas 60000 para treinamento e 10000 para teste, previamente definidas, com o intuito de obter as melhores combinações de parâmetros C e γ, por meio do algoritmo genético. Em seguida, verificou-se a eficiência da SVM com cada combinação de parâmetros variando as amostras de treinamento e teste aleatoriamente nesses experimentos. Eles diferenciaram-se em quantidade de amostras utilizadas para testes, sendo 10%, 20% e 30% dos 70000 caracteres.

O cromossomo utilizado para representar as características contínuas dos indivíduos possui duas variáveis , C e γ. O máximo de gerações utilizado foram de 20, enquanto a população inicial foi formada por 20 indivíduos gerados aleatoriamente, com os parâmetros C e γ dentro dos intervalos [0,100] e [0,0.1], respectivamente, por terem apresentados os melhores indivíduos ao longo das simulações, fazendo com que o AG convirja mais rapidamente. A função de avaliação utilizada calcula o percentual de acertos do classificador SVM com os parâmetros selecionados. Já o método de seleção abordado foi o de torneio, selecionando uma quantidade de indivíduos igual ao da população inicial. A reprodução foi realizada por meio do *Blend-Crossover* e mutação aplicada foi a . Os critérios de parada utilizados foram a obtenção do máximo de gerações ou a não obtenção de um novo indivíduo melhor que os existentes por algumas gerações.

Foram realizadas diversas simulações nos três experimentos com o sistema HCR híbrido, utilizando uma SVM e um Algoritmo Genético para estimar os parâmetros da SVM. A Tabela 4 apresenta as melhores combinações dos parâmetros, C e γ, encontrados pelo Algoritmo Genético na primeira coluna e as acurácias médias obtidas nos experimentos utilizando 10%, 20% e 30% das 70000 amostras para teste.

Os melhores parâmetros C e γ encontrados pelo algoritmo genético apresentaram elevados indíces de HCR nos três experimentos realizados, variando a quantidade de amostras utilizadas para testes. As simulações de número 1, 2 e 3 apresentadas na Tabela 4, obtiveram percentuais de acertos superiores em relação as demais, em praticamente todos os experimentos.

Número	C,γ	Exp. 1 (10%)	Exp. 2 (20%)	Exp. 3 (30%)
1	31.4503, 0.0868	97,79%	97,76%	97,65%
2	33.609, 0.0857	97,81%	97,75%	97,65%
3	32.1846, 0.0868	97,79%	97,76%	97,66%
4	30.9, 1.0451	96,54%	96,85%	96,43%
5	26, 0.0919	97,77%	97,68%	97,65%
6	58, 0.0578	97,75%	97,61%	97,60%
7	56.04, 0.059	97,77%	97,62%	97,60%

Tabela 1. Melhores parâmetros encontrados pelo Algoritmo Genético.

5. Conclusões e Perspectivas Futuras

Os resultados obtidos com a utilização de um algoritmo genético para o HCR mostraram-se superiores em relação a trabalhos que não utilizaram técnicas de otimização, como é o

caso dos trabalhos de [Kumar et al. 2012], que atingiu uma acurácia máxima de 94,8%, e [Rodrigues et al. 2001] que alcançou 94,64%.

Com o objetivo de otimizar e melhorar a precisão no HCR, outras técnicas de extração de características serão abordadas no sistema HCR, assim como também utilizar o algoritmo genético para selecionar características a serem processadas no processo de classificação.

Referências

Bonesso, D. (2013). Estimação dos Parâmetros do Kernel em um Classificador SVM na Classificação de Imagens Hiperespectrais em uma Abordagem Multiclasse. Master's thesis, Universidade Federal do Rio Grande do Sul Centro Estadual de Sensoriamento Remoto e Meterologia Programa de Pós-Graduação em Sensoriameno Remoto, Porto Alegre.

Coelho, G. V. V. (2013). Seleção de Características usando Algoritmos Genéticos para Classificação de Imagens de Textos em Manuscritos e Impressos. Master's thesis, Universidade Federal de Pernambuco, Recife.

Gonzalez, R. C. and Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kumar, P., Sharma, N., and Rana, A. (2012). Handwritten character recognition using different kernel based svm classifier and mlp neural network (a comparison). *International Journal of Computer Applications*, 53(11).

LeCun, Y., Cortes, C., and Burges, C. J. (1998). The mnist database of handwritten digits. Disponível em: <https://goo.gl/LU1HuQ>. Acesso em: 20 de abril de 2017.

Linden, R. (2012). *Algoritmos Genéticos*. Editora Ciência Moderna, 3rd edition.

Mori, S., Suen, C. Y., and Yamamoto, K. (1995). Document image analysis. chapter Historical Review of OCR Research and Development, pages 244–273. IEEE Computer Society Press, Los Alamitos, CA, USA.

Pradeep, J., Srinivasan, E., and Himavathi, S. (2012). Neural network based recognition system integrating feature extraction and classification for english handwritten. *International journal of Engineering*, 25(2):99–106.

Rodrigues, R. J., Silva, E., and Thomé, A. C. G. (2001). Feature extraction using contour projection. Disponvel em: <https://goo.gl/M3YjoM>. Acesso em: 25 de julho de 2017.

Trier, Ø. D., Jain, A. K., and Taxt, T. (1996). Feature extraction methods for character recognition - a survey. *Pattern Recognition (PR)*, 29(4):641–662.

Yerra, N., Varanasi, R., Adapaka, H., Surumalla, H., and Dantha, J. (2017). Recognition of handwritten characters using svm. *International Journal of Innovate Research in Science and Engineering*, 4(3).