Comparação de classificadores para o reconhecimento de caracteres de documentos manuscritos do português do seculo XIX

Sergio D Navarro López

***Resumo*—Este documento contem a proposta de trabalho de pesquisa da disciplina do programa de Pós-Graduação em Engenharia Elétrica UFBA, Inteligência computacional.**

***Index Terms*—Inteligência, Computacional, Support Vectors Machine, OCR, MLP, Redes Neurais**

I. Introdução

PESQUISADORES da área das ciências humanas interessados em realizar pesquisas sobre a historia do Brasil utilizando manuscritos do seculo XIX encontram grandes dificuldades no que tange à leitura e transcrição desses documentos. A escazes de profissionais especializados nessa atividade faz com que o pesquisador seja obrigado a realizar um grande investimento financeiro para realizar essa tarefa. Nesse sentido a pesquisa descrita neste documento visa qualificar uma metodologia da área da inteligência computacional para o reconhecimento de caracteres manuscritos do português do período em estudo.

II. Origem dos datos

Como resultado do projeto desenvolvido pela Universidade Estadual de Feira de Santana - UEFS: Cativos às portas do sertão: fontes para a história da escravidão e das populações negras em Feira de Santana e Região (1830-1885), foi criado um acervo de documentos manuscritos digitalizados de acesso publico em formato jpg de alta qualidade [1].

Desses arquivo de imagem em alta resolução e de outras fontes existentes e com ajuda de um profissional na transcrição de manuscritos, serão extraídas mostras de cada carácter do alfabeto português.

III. Etapas

O fluxo proposto para avaliação da técnica de reconhecimento de caracteres esta composta das seguintes processos, tomando como base o proposto por [2], [3]:

- Pré-processamento
- Segmentação
- Extração
- Classificação.

IV. Pré-processamento

As imagens dos documentos que foram digitalizados contem níveis apreciáveis de ruido. O pré-processamento é realizado para diminuir esse níveis de interferência. Nessa etapa serão usadas API de bibliotecas para tratamentos e transformação de imagens para realizar filtrado, dilatação, erosão, redimensionamento, etc [3].

V. Segmentação

Em esta etapa uma sequência de caracteres de uma imagem é particionada em varias imagens de caracteres individuais. Também é necessário normalizar o tamanho da imagem para ajustá-lo ao template de classificação e determinar a posição do carácter na imagem digital [3]. Podem-se apresentar algumas dificuldades de segmentação devido a natureza do manuscrita dos documentos da época.

VI. Extração

O objetivo da extração e capturar as características essenciais dos simbolos. Estas características são a principal entrada para os processos de classificação e reconhecimento. A eficiência deste processo impacta a taxa de reconhecimento [4] do algoritmo de classificação.

Existem vários métodos avaliados anteriormente e no presente trabalho serão trabalhados os seguintes mecanismos de extração:

- Multifactor Dimensionality Reduction
- Principal Components Analysis

VII. Classificação

Para avaliação dos métodos de classificação sera implementado SVM (Support Vector Machine) e MLP (Multi-layer Perceptron) [5] que permitem um implementação simples em linguagem de programação como C, C++ ou Python, como também em ferramentas de MatLab ou Octave [6].

A. SVM - Support Vector Machine

É uma técnica de classificação de dados. As tarefas de classificação comumente requer separação de dados para ser usados como dados de teste e dados de treinamento. Cada instancia no conjunto de dados de treinamento contem um valor alvo e vários atributos. O objetivo principal da SVM é produzir um modelo baseado nos dados de treinamento que predize los valores alvo dos dados de prova usando somente seus atributos.

Se propõe acompanhar o seguinte fluxo [7]:

- Transformar os dados a ser analisados a um formato compatível de entrada para o SVM
- Realizar normalização dos dados ou reduzir a escala dos mesmos
- Considerar o kernel RBF e encontrar os melhores parâmetros C e γ

- Usar os parâmetros C e γ para treinar o conjunto de treinamento
- Testar.

B. MLP - Multi Layer Perceptron

Outra técnica para realizar a classificação de dados e treinamento é o algoritmo de aprendizado Multi-layer Perceptron, como esta descrito em [8]. Os pesos sinápticos e valores threshold são atualizados da forma em que as tarefas de classificação e reconhecimento sejam realizadas eficientemente. Em este tipo de rede neural, na entrada existe uma camada de dados que e gerada na etapa de extração de características e na saída uma camada que representa um vector. Inicialmente sera usada uma rede neural de 3 camadas. Os parâmetros da rede neural serão inferidos dos parâmetros resultantes da etapa de segmentação e extração.

REFERÊNCIAS

[1] (2008) Cativos às portas do sertão. [Online]. Available: http://aquarios.uefs.br:8081/cativosdosertao/index.html

[2] C. Halder, J. Paul, and K. Roy, "Comparison of the classifiers in bangla handwritten numeral recognition," in *Radar, Communication and Computing (ICRCC), 2012 International Conference on*. IEEE, 2012, pp. 272–276.

[3] D. Singh, M. A. Khan, A. Bansal, and N. Bansal, "An application of svm in character recognition with chain code," in *Communication, Control and Intelligent Systems (CCIS), 2015*. IEEE, 2015, pp. 167–171.

[4] J. Pradeep, E. Srinivasan, and S. Himavathi, "Diagonal based feature extraction for handwritten character recognition system using neural network," in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, vol. 4. IEEE, 2011, pp. 364–368.

[5] C. C. Gohel, M. M. Goswami, and V. K. Prajapati, "On-line handwritten gujarati character recognition using low level stroke," in *Image Information Processing (ICIIP), 2015 Third International Conference on*. IEEE, 2015, pp. 130–134.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[7] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

[8] T. Alam *et al.*, "An approach to empirical optical character recognition paradigm using multi-layer perceptorn neural network," in *Computer and Information Technology (ICCIT), 2015 18th International Conference on*. IEEE, 2015, pp. 132–137.