Análise de classificadores

Aryane Ast dos Santos
Departamento de Informática
Universidade Federal do Paraná
Email: aras10@inf.ufpr.br

***Resumo*—Abstract goes here**

I. Introdução

Um problema de classificação consiste em definir um rótulo ou classe para um elemento a partir de um conjunto de elemento com rótulos definidos. É um problema de aprendizagem supervisionada, cujo objetivo é realizar inferências a partir de um conjunto de dados rotulados, em oposição à aprendizagem não-supervisionada.

Este relatório se propõe a apresentar resultados obtidos com os classificadores*K Nearest Neighbors* (KNN), *Naive Bayes*, Árvores de Decisão e *Support Vector Machines* (SVM) num problema de classificação de imagens, cuja base rotulado possui 1901 imagens em 9 classes diferentes. Os algoritmos de classificação não utilizam as imagens brutas, de forma que é necessário converter as imagens do formato JPG para vetores de características que os algoritmos de classificação possam utilizar.

Após extraído o vetor de características a partir das imagens, foram realizadas as execuções dos classificadores KNN, Naive Bayes, Árvores de Decisão e SVM. As implementações dos algoritmos mencionados são da biblioteca Scikit Learn (ref).

Nas seções a seguir são apresentados maiores detalhes da representação, algoritmos utilizados, métricas para comparação e desempenho. São comparados também o desempenho de estratégias de combinação de classificadores e *ensembles*.

II. Representação dos dados

Para cada uma das imagens disponibilizadas para classificação, é realizada uma extração de características, que resulta num vetor com as características Para a extração dos vetores de caracteristicas, foram utilizados os algoritmos *Local Binary Patterns* (LBP) e *Grey-Level Co-Occurrence Matrix* (GLCM), o que resultou em vetor contendo 24 características, além da classe ao final da linha.

A. Local Binary Patterns

Breve explicação. Método uniforme, raio=2, n_point ou vizinhos = 16, implementação do scikit learn.

B. Grey-Level Co-Occurrence Matrix

Breve explicação.

Características utilizadas: correlação, dissimilarity, contrast, homogeneity, energy, ASM.

III. Classificação

A partir dos vetores de características, é possível executar os algoritmos de classificação. Como temos apenas uma base de dados, se a utilizarmos inteira para treinar os algoritmos e após isso, testar se a classificação é feita corretamente com essa mesma base, ocorrerá algo chamado de *overfitting*, que é quando a base é muito especializada e acerta predições para um conjunto de dados conhecido, mas para dados desconhecidos costuma errar. Para fugir dessa situação, é boa prática separar a base em treinamento e validação.

Entretanto, ao separar a base em treinamento e validação, reduz-se muito a quantidade de dados dos quais se aprende (dados treinamento). Para evitar tal situação, se faz uso de uma técnica chamada validação cruzada ou *cross-validation*, onde se separa ...

Neste trabalho, para a validação cruzada são utilizados os métodos ShuffleSplit e cross_val_score do módulo model_selection da biblioteca SciKit Learn. Dessa forma, a base é dividida 10 vezes em treinamento e validação nas proporções de 0.6 e 0.4 respectivamente.

A. KNN

Breve explicação.
Resultados.

B. Naive Bayes

Breve explicação.
Resultados.

C. Árvores de decisão

Breve explicação
Resultados.

D. SVM

Breve explicação.
Resultados.

E.

IV. Considerações Finais

Todo o código utilizado no projeto, inclusos ..., pode ser encontrado num repositório Git hospedado no GitHub ....