Redução de *Template* aplicado a *K-Nearest Neighbor* e Floresta de Caminhos Ótimos em problemas de Reconhecimento de Padrões

Francisco Samuel Lemke

Paulo Alberto Thiry Neto

Trabalho de Conclusão de Curso
a
FACET (Faculdade de Ciências Exatas e Tecnológicas)
da
Universidade Federal da Grande Dourados
para
obtenção do título de graduação
em
Análise de Sistemas

Programa: Graduação em Análise de Sistemas - FACET (UFGD)

Orientador: Prof. Msc. Willian Paraguassu Amorim

Dourados, fevereiro de 2013

UNIVERSIDADE FEDERAL
DA GRANDE DOURADOS

Redução de *Template* aplicado a *K-Nearest Neighbor* e Floresta de Caminhos Ótimos em problemas de Reconhecimento de Padrões

Este exemplar corresponde ao projeto final de
Graduação em Análise de Sistemas, elaborado pelos
acadêmicos Francisco Samuel Lemke e Paulo Alberto Thiry Neto
sob orientação do Prof. Msc. Willian P. Amorim.

Francisco Samuel Lemke
Paulo Alberto Thiry Neto

Redução de *Template* aplicado a *K-Nearest Neighbor* e Floresta de Caminhos Ótimos em problemas de Reconhecimento de Padrões

Trabalho de Conclusão de Curso aprovado como requisito parcial para obtenção do título de Bacharel em Análise de Sistemas na Universidade Federal da Grande Dourados, pela comissão formada por:

Orientador: Prof. Msc. Willian Paraguassu Amorim

Prof. Dr. Adailton José Alves da Cruz

Prof. Msc. Murilo Táparo

Dourados, 09 de abril de 2013

Resumo

Lemke, F. L. Neto, P. A. T. **Redução de Template aplicado a K-Nearest Neighbor e Floresta de Caminhos Ótimos em problemas de Reconhecimento de Padrões.** (Projeto de Graduação) - FACET (Faculdade de Cinências Exatas e Tecnológicas), Universidade Federal da Grande Dourados, Dourados, 2012.

Um dos algoritmos clássicos da área de aprendizagem e classificação de dados é o *KNN* (*K Nearest Neighbors*). O algoritmo *KNN* é de aprendizado supervisionado, sendo o resultado da consulta de uma nova instância é classificado com base na classe majoritária dos K vizinhos mais próximos. O *KNN* apresenta algumas desvantagens, como a cada nova consulta será necessário calcular a distância para todas as amostras de treinamento para chegar a uma determinada classificação, exigindo um elevado tempo de processamento. Outra abordagem que vem sendo utilizada na classificação de padrões é o *OPF* (Floresta de Caminhos Ótimos). *OPF* é um *framework* para classificação de dados supervisionados e não supervisionados, particularmente eficiente na classificação de imagens, que reduz problemas de classificação de padrões para um problema de particionamento dos vértices de um grafo. Um dos possíveis problemas que pode ser encontrado no uso da técnica *OPF*, é a classificação de grandes volumes de dados. Para a fase de treinamento é necessário calcular o grafo completo, ou seja, a distância entre todas as amostras, o que em uma base de treinamento muito grande usará grande tempo de processamento. O método *Template Reduction* (*TR*), proposto neste trabalho, tem como objetivo obter um modelo menor, um subconjunto extraído do conjunto de treino. A ideia principal do *TR* é encontrar as regiões de fronteira entre as diferentes classes do conjunto de treino, e descartar os protótipos que estão longe desta região. Neste trabalho, avaliamos o uso do *TR*, e após a aplicação ao *KNN* e *OPF*. Avaliamos, pelos resultados, que a proposta apresentada *TR*, obteve resultados satisfatórios, mesmo não conseguindo os melhores resultados de taxa de classificação, ainda se mostra eficiente, diminuindo significativamente o volume de dados, e acelerando o processo de aprendizagem e classificação.
Palavras-chave: Reconhecimento de Padrões, Aprendizado de Máquina, Template Reduction, KNN, Floresta de Caminhos Ótimos.

Abstract

Lemke, F. L. Neto, P. A. T. **Template Reduction applied to K-Nearest Neighbor and Optimum-Path Forest in problems Pattern Recognition.** (Projeto de Graduação) - FACET (Faculdade de Cinências Exatas e Tecnológicas), Universidade Federal da Grande Dourados, Dourados, 2012.

An of the classic algorithms in the area of learning and classification of data is the KNN (K Nearest Neighbors). The KNN algorithm is supervised learning, where the consultation result of a new instance is classified based on the dominant class of K nearest neighbors. The KNN has some disadvantages, as each new consultation will need to calculate the distance to all training samples to reach a certain ranking, requiring a high processing time. Another approach that has been used in pattern classification is the OPF (Optimum Path-Forest). OPF is a framework for data classification, supervised and unsupervised, particularly effective in image classification, which reduces problems of pattern classification to a problem of partitioning the vertices of a graph. One of the possible problems that may be encountered in the use of the technique OPF is the classification of large volumes of data. For the training phase is necessary to calculate the complete graph, in other words, the distance between all samples, which in a training base too large will cause a high processing time. The method Template Reduction (TR), the present paper aims to obtain a smaller model, a subset extracted from the training set. The main idea is to find the TR border regions between different classes of the training set, and discard prototypes that are far from this region. In this study, we evaluated the use of TR, and after applying the KNN and OPF. We evaluate the results, the proposal TR, achieved satisfactory results, while not getting the best results of classification rate, yet is efficient, significantly reducing the volume of data, and accelerating the learning process and classification.
Keywords: Pattern Recognition, Machine Learning, Template Reduction, KNN, Optimum-Path Forest.