k-NN com Algoritmo Genético: Uma comparação

Marco Cezar Moreira de Mattos[1], **Rômulo Manciola Meloca**[1]

[1]DACOM – Universidade Tecnológica Federal do Paraná (UTFPR)
Caixa Postal 271 – 87301-899 – Campo Mourão – PR – Brazil

{marco.cmm,rmeloca}@gmail.com

Resumo. *Relata a aplicação de algoritmo Genético para melhorar o desempenho do algoritmo k-NN.*

1. O Problema

Para um humano reconhecer quais dígitos foram manuscritos por outro humano, embora algumas caligrafias sejam difíceis de lidar, esta é uma tarefa trivial, pois, em geral, o cérebro humano consiga facilmente decifrar um número, seja pelo contexto, seja pelo costume ou pela especialidade de quem decifra. São vários os atributos que um número caligrafado pode possuir, tais como: tremulação, inclinação da caneta, pressão exercida sobre o papel, velocidade de escrita e formato.

Já para uma máquina, esta tarefa não possui a mesma trivialidade. É claro que comparar se uma imagem é idêntica a outra é fácil, no entanto, sabemos que não escrevemos de maneira idêntica todos os números, já que dos número advém, inclusive, emoções do escritor. Uma máquina deveria de conhecer todos os possíveis tipos de escrita dos números e ainda assim o algoritmo de busca seria lento, dado a variedade de tipos. Cabe portanto, algoritmos de busca local ou algoritmos que aprendam, uma vez que estes não olham para o todo, mas procuram adaptar-se a fim de obter um resultado para determinada entrada.

Deste modo, dado as características extraídas sobre cada instância aplicou-se, neste trabalho, dois algoritmos um para selecionar as melhores características e outro para classificá-las e observar as taxas de acerto. Para selecionar as melhores características utilizou-se o algoritmo genético, um tipo de algoritmo evolucionário, que é incapaz de aprender (portanto um algoritmo de busca) classificado como uma busca local. Para solução escolheu-se, dentre os algoritmos com aprendizagem supervisionada (aqueles que operam sobre um conjunto de aprendizagem com as respostas) o k-Nearest Neighbor (k-NN).

2. Solução

Para o conjunto de treinamento, foram dadas 1000 instâncias com 132 características. Das instâncias, cada uma pertencia a uma das 10 classes encontradas no conjunto (evidente, pois cada instância correspondia a um número de 0 à 9). Para cada instância encontrava-se disponibilizado a classe (em outro arquivo) a que, de fato, a instância pertencia, uma vez que foi utilizado um algoritmo de aprendizagem supervisionada, que precisa saber, de antemão, a resposta de cada instância.

Para o conjunto de teste, também, foram dadas 1000 instâncias com 132 características cada, bem como a classe de cada instância (em outro arquivo).

Para melhorar os resultados obtidos pelo k-NN é utilizado o algoritmo genético para aprimorar a população (que é um sub-conjunto do conjunto de treino). A população especializada pelo algoritmo genético serviu de entrada para o k-NN poder classificar o conjunto de teste. Sua função foi a de obter as melhores características a serem tomadas para o algoritmo k-NN.

Utilizou-se a linguagem de programação MATLAB, por possuir muitas funcionalidades prontas, como o k-NN e métodos prontos para manipular matrizes.

2.1. Algoritmo Genético

O algoritmo Genético é um algoritmo de busca local que toma uma população, aleatória ou não, e faz sobre ela operações genéticas (como acontece com o DNA) de cruzamento e mutação, que também podem ser escolhidas de maneira aleatória. Apropria-se da ideia de evolução das espécies, de Darwin, para especializar, dentre a população, sempre os melhores resultados de modo a convergir para o máximo global. Sua aleatoriedade que, justamente, permite resultados muito bons por não limitar-se a máximos locais.

2.2. k-NN

O algoritmo k-NN é um algoritmo de aprendizagem supervisionada que toma um conjunto de treinamento com suas respectivas respostas e classifica as instâncias de acordo com a maioria dos k (ímpares para desempate) vizinhos mais próximos da instância dada. Para calcular a distância dos vizinhos existem várias fórmulas. A fórmula mais comum para a métrica é a distância euclidiana.

3. Resultados

Os resultados obtidos encontram-se descritos nas seções que seguem:

3.1. Base: Apenas o k-NN

Inicialmente realizou-se um teste apenas com o k-NN, sem a intervenção do algoritmo Genético. Assim, obteve-se uma base de 89,99% de acerto.

3.2. k-NN com Algoritmo Genético

Em seguida, executou-se o algoritmo com os seguintes parâmetros:

- Caracteristicas: 132
- Tamanho da população: 10
- Quantidade gerações: 100

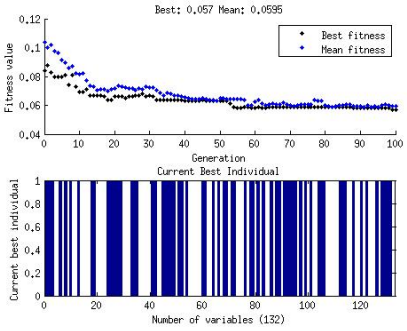

Figura 1. Execução 1

Obteve-se pelo algoritmo o resultado de 0,057. Assim, temos que a taxa de acerto foi 94.3%.

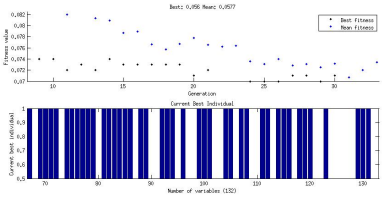

Figura 2. Execução 2

Obteve-se pelo algoritmo o resultado de 0,059. Assim, temos que a taxa de acerto foi 94.1%.

Em seguida, executou-se o algoritmo com os seguintes parâmetros:

- Caracteristicas: 132
- Tamanho da população: 10
- Quantidade gerações: 20

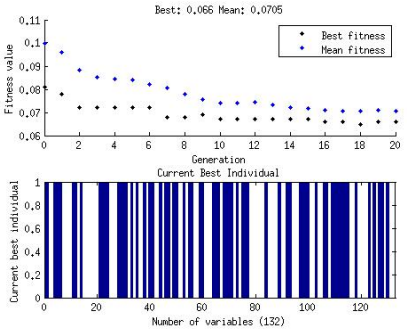

Figura 3. Execução 3

Obteve-se pelo algoritmo o resultado de 0,066. Assim, temos que a taxa de acerto foi 93.4%.

Na sequência, executou-se o algoritmo com os seguintes parâmetros:

- Caracteristicas: 132
- Tamanho da população: 50
- Quantidade gerações: 200

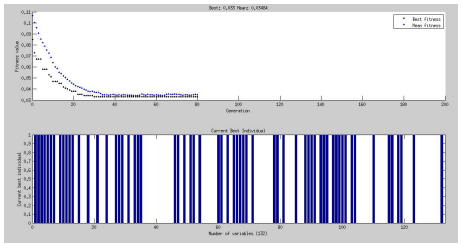

Figura 4. Execução 4

Obteve-se pelo algoritmo o resultado de 0,033. Assim, temos que a taxa de acerto foi 96.7%.

4. Considerações Finais

Como sugerido por CAVALCANTI et al. o que mais impacta nas taxas de acerto desse tipo de problema são a qualidade das características obtidas. Pode-se observar, portanto, que, embora as taxas de acerto utilizando o k-NN "puro", isto é, com as 132 características coletadas tenham sido satisfatórias, ao utilizar o algoritmo genético para pinçar as melhores características a serem tomadas, as taxas sofreram aumento considerável, o que justamente comprova o trabalho de Cavalcanti.

Deste modo, ao decorrer do experimento foi possível notar que a utilização do algoritmo genético para aprimorar o conjunto de treinamento, que é a entrada do algoritmo k-NN, sempre impactou em melhora significativa das taxas de acerto.

Notou-se ainda que a medida em que aumentava-se o número de instâncias tomadas como população, embora o tempo de execão do algoritmo aumentava exponencialmente, ou até mesmo quando acrescido a quantidade de gerações utilizadas para o algoritmo genético, o resultado tinham suas taxas de acerto melhoradas.

5. Referências

CRUZ, Rafael M. O., CALVALCANTI, George D. C., REN, Tsang I. **Análise de Técnicas de Extração de Características para o Reconhecimento de Dígitos Manuscritos**.