Uma proposta de versão contextual para o classificador de vizinhos mais próximos

Flavia de Toledo Martins Bedê[1], Luciano Vieira Dutra[2], Sandra Sandri[3]

[1]Programa de Mestrado ou Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Divisao de Processamento de Imagens – DPI
Instituto Nacional de Pesquisas Espaciais – INPE

[3]Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

{flavinha,dutra}@dpi.inpe.br, sandri.at.lac.inpe.br@gmail.com

Resumo. *A versão contextual do método K-NN, identificada como MS-NN (Multi Space – Nearest Neighborhood), incorpora em sua solução um parâmetro k para cada tipo de dado (espectral, espacial e temporal, por exemplo). Além disso, outro diferencial do método proposto é poder usar, ou não, diferentes tipos de distâncias para cada tipo de espaço.*

Abstract. *The K-NN method contextual version, identified as MS-NN (Multi Space - Nearest Neighborhood), adds a parameter k for each type of data (spectral, spatial and temporal, for example) in the solution. In addition, another difference of the proposed method is able to use, or not, different types of distances for each type of space.*

Palavras-chave: *Classificação, K-NN, modelagem.*

1. Introdução

O K-NN é uma técnica de classificação de padrões que consiste em atribuir uma classe a um elemento desconhecido usando a classe da maioria de seus vizinhos mais próximos, segundo uma determinada distância (no espaço de atributos). A sua versão estendida, identificada como Multi Space – Nearest Neighborhood (MS-NN), irá basicamente incorporar na construção do modelo a utilização de múltiplos espaços de atributos semanticamente distintos. Para cada espaço pode ser utilizado um número de vizinho diferente e uma distância diferente. A construção do modelo MS-NN baseia-se na procura dos vizinhos para cada distância utilizada. Por exemplo, MS-NN com 3 espaços, pode-se considerar 3 distâncias distintas (l_1, l_2, l_3), ou seja, k_1 vizinhos mais próximos, considerando uma distância espectral (exemplo distância euclidiana), a k_2 vizinhos mais próximos, considerando uma distância geográfica (distância real entre as sedes de duas cidades por ex.) e a k_3 vizinhos mais próximos considerando uma distância no tempo.

O objetivo deste trabalho é propor o modelo MS_NN que considera dois diferentes tipos de distâncias, particularmente a distância geográfica Assim, o modelo contextual proposto visa incorporar a informação do espaço na classificação.

2. MS-NN – um modelo de vizinhos mais próximos para *l* distâncias distintas

Nesta seção será apresentada a definição do modelo K-NN além de uma breve explicação sobre o modelo MS-NN proposto.

Em reconhecimento de padrões, o algoritmo K-NN é um dos mais simples de todos os algoritmos de aprendizado de máquina. Baseado na analogia, um objeto é classificado pelo voto da maioria de seus vizinhos. Este processo de classificação pode ser computacionalmente exaustivo se considerado um conjunto com muitos dados. Por isso, a grande desvantagem é o tempo de computação para a obtenção dos *k* vizinhos mais próximos. Então a maioria dos estudos envolvendo K-NN tem o objetivo de aumentar a eficiência computacional e reduzir a taxa de erro de generalização deste método (Bishop, 2007; Michie e Spiegelhalter, 1994; Webb, 2002).

O K-NN possui apenas um parâmetro livre (o número de *k* vizinhos) que é controlado pelo usuário com o objetivo de obter uma melhor classificação. O melhor valor de *k* pode ser determinado experimentalmente. Começa-se com $k = 1$, e utiliza-se um conjunto de testes, para estimar a taxa de erro do classificador. Para cada *k*, classificam-se as tuplas do conjunto de testes e verifica-se quantas tuplas foram classificadas corretamente. O valor de *k* que apresentar a menor taxa de erro será o escolhido. Normalmente, os valores de *k* escolhidos são 1, 2, 3 ou $\sqrt{n}$, onde *n* é o tamanho da base de treinamento (Bishop, 2007; Webb, 2002).

Basicamente, uma base de dados de treinamento composta por um conjunto de tuplas $\{a_1, ..., a_n, cl\}$, onde *cl* é a classe à qual pertencem as tuplas $\{a_1, ..., a_n\}$, é usada para classificar um novo caso c_0 (representado como $c_0 = (a_1(c_0),..., a_n(c_0))$). A classificação é realizada da seguinte maneira (Theodoridis e Koutroumbas, 2006):

- Inicialmente estabelece-se um valor para k (geralmente se determina um valor impar para *k,* de forma que este valor não seja múltiplo do número total de classes);
- Calculam-se as distâncias *x* do caso c_0 às todos os casos de treinamento;
- Identifica-se os *k* vizinhos mais próximos, independentemente do rótulo de classe;
- Dentro das *k* casos identificados, identificar o número de casos que pertencem a cada classe;
- Classifica-se o caso c_0 associando-se a ele a classe mais frequente, ou seja, a classe que a maioria das *k* casos pertence.

Pode ser útil atribuir pesos às contribuições dos vizinhos, de modo que os vizinhos mais próximos contribuem mais para a média do que os mais distantes. O mais comum é atribuir a cada vizinho o peso de *1/x*. Também é possível atribuir um peso relativo à importância de cada atributo (Bishop, 2007; Webb, 2002).

Como dito anteriormente, o processo de classificação pode ser computacionalmente exaustivo. A maioria das pesquisas em K-NN se concentra na

tentativa de aumentar a velocidade para calcular o vizinho mais próximo (Fukunaga e Narendra, 1975; Gates, 1972).

2.1. Principais elementos do modelo proposto

Como o K-NN possui apenas um parâmetro livre, *k*, a proposta do MS-NN é inserir mais um parâmetro, que define quantos espaços quer se trabalhar e para cada espaço, uma distância diferente, ou não, pode ser escolhida. Assim, o usuário poderá controlar mais de um parâmetro, dependendo dos dados e da aplicação.

No modelo MS-NN, onde MS = (k_1, k_2, ..., k_l) a definição de quantos vizinhos devem ser usados para cada espaço (espectral, espacial e temporal, por exemplo) pode ser feita da mesma maneira que se define o valor *k*, i.e., experimentalmente. Para melhor compreensão, os passos a seguir mostra como seria a classificação usando dois espaços de atributos, e duas distâncias distintas. Como exemplo, para o primeiro espaço é usado a distância euclidiana e para o segundo, uma distância geográfica. Utiliza-se 2 espaços de atributos.

1. Para o primeiro espaço:
   - Estabelece-se um valor para o número de vizinhos (*k*)
   - Calcula-se as distâncias do caso c_0 à todos os casos de treinamento.
   - Identifica-se os rótulos dos *k* casos mais próximos;
2. Para o segundo espaço:
   - Identifica-se os vizinhos de fronteira de c_0;
   - Estabelece-se um número *x* para soma dos pesos (esse número será divido pelo total de vizinhos de c_0);
   - Atribui-se o peso (*x*/ total de vizinhos de c_0) para os rótulos dos casos vizinhos;
3. Dentre os *k* rótulos identificados no passo 2 e os rótulos com peso do passo 3, identifica-se o número de casos que pertencem a mesma classe.
4. Classifica-se o caso c_0 associando-se a ele a classe mais frequente.

3. Resultados esperados

Pretende-se usar o algoritmo MS-NN proposto para classificar a prevalência da esquistossomose em municípios de Minas Gerais, usando variáveis derivadas de Sensoriamento Remoto (SR), climáticas, e socioeconômicas. Este algoritmo poderá ser aplicado a qualquer outro problema que tenha ou não dependência espacial.

Referencias

Bishop, C. M. Pattern Recognition and Machine Learning Springer, 2007. 738 p.

Fukunaga, K.; Narendra, P. M. A branch and bound algorithm for computing nearest neighbours. IEEE Trans. Comput, p. 917–922, 1975.

Gates, G. W. The reducednearestneighbour rule. IEEETransactions on Information-Theory, p. 431, 1972.

Michie, D.; Spiegelhalter, D. J. Machine Learning, Neural and Statistical Classification Prentice Hall, 1994. 289 p.

Theodoridis, S.; Koutroumbas, K. Pattern recognition Academic Press, 2006. 885 p.

Webb, A. R. Statistical Pattern Recognition. 2nd. London, U.K.: Wiley, 2002. 496 p.