Algoritmo dos K-Vizinhos Mais Próximos Baseado em Grafo

Gustavo F. C. de Castro, Renato Tinós

Departamento de Computação e Matemática
Faculdade de Filosofia Ciências e Letras de Ribeirão Preto (FFCLRP)
Universidade de São Paulo (USP) – Ribeirão Preto, SP – Brasil

gus.castro@usp.br, rtinos@ffclrp.usp.br

Objetivos

O objetivo principal deste trabalho é a investigação de estratégias eficientes utilizando o Grafo de Interações Nk [6] para retornar os *K* exemplos mais próximos da base de treinamento que serão utilizados para classificar um novo exemplo no algoritmo dos K-Vizinhos Mais Próximos (KNN - K-Nearest Neighbors). Os objetivos secundários são: explorar as propriedades derivadas da fusão do KNN e do Grafo de Interações Nk; modificar a forma como o Grafo Nk é construído; comparar as diferenças entre os algoritmos ao variar separadamente os parâmetros *K*, que define o número de arestas do grafo, e *α*, que define a razão entre arestas definidas pela densidade espacial e pela distância Euclidiana; comparar as variantes do KNN modificado com o KNN original, utilizando diferentes valores para parâmetros, em bases de dados [2] [4] com propriedades distintas.

Métodos e Procedimentos

O KNN proposto utiliza do Grafo de Interações Nk para retornar os *K* vizinhos mais próximos de um novo exemplo (a ser classificado) *x*. O KNN é um algoritmo simples e eficiente [3] de aprendizado supervisionado [5] no qual um novo exemplo *x* é classificado examinando-se a classe majoritária entre os *K* exemplos da base de treinamento mais próximos a *x* [1]. No KNN original, os *K* exemplos mais próximos são definidos pela menor distância Euclidiana a *x*.

O Grafo de Interações Nk é um grafo direcionado com *N* vértices de grau de saída *k* [6]. Cada vértice é associado a um exemplo da base treinamento, e cada aresta ($\boldsymbol{v}_j$, $\boldsymbol{v}_i$) do grafo indica a relação do *j*-ésimo exemplo com *i*-ésimo exemplo. Neste projeto, foi proposto um parâmetro *α* que especifica a razão das arestas definidas por densidade. Sabe-se que $A = \lfloor \alpha k \rfloor$, onde *A* é um inteiro que representa a quantidade de arestas de saída definidas por densidade espacial, para cada vértice. No grafo Nk original, *α* sempre leva a *A*=1 e, portanto, sempre haverá um auto-loop, uma aresta definida por densidade e *k*-2 arestas definidas por distância, para $k \geq 2$. Para $k = 1$, sempre haverá apenas o auto-loop (*A*=0). Aqui é proposto o uso de *α* tal que *A*≠1, mudando o número de arestas definidas por ambas métricas. Para classificar um novo exemplo *x*, o vértice $\boldsymbol{v}_x$ cujo exemplo é o mais próximo a *x* é escolhido para representá-lo no Grafo de Interações Nk. Duas variações do algoritmo são propostas: na primeira variação, KNN modificado tipo A, a lista de adjacência de $\boldsymbol{v}_x$ é obtida utilizando densidade e distancia. Os *k*=*K* exemplos da lista de adjacências são utilizados na classificação; A segunda variação, KNN modificado tipo B, consiste em salvar o vértice $\boldsymbol{v}_j$, da lista de adjacência de $\boldsymbol{v}_x$, cujo exemplo $\boldsymbol{y}_i$ é o mais próximo a *x*, ignorando o auto-loop. Então a operação $\boldsymbol{v}_x = \boldsymbol{v}_j$ é realizada, e a etapa anterior é repetida *k* vezes. A lista de *k* vértices $\boldsymbol{v}_j$ gerada por este processo é utilizada na classificação.

Resultados

Os algoritmos propostos são comparados, quanto à precisão (ACC) e número de erros (n), ao KNN original em dois experimentos. Estes foram projetados para testar os efeitos da alteração de *K* (Experiência 1) e α (Experiência 2), utilizando validação cruzada de dez folds. Nas tabelas, os resultados destacados em cinza escuro são os melhores resultados gerais da respectiva base de dados, enquanto os resultados destacados em cinza claro

representam os melhores resultados de sua respectiva linha (para um dado valor *K* ou *α*)

Compound -> α = 0.1							
K	A		B		KNN		Total
	n	*ACC*	*n*	*ACC*	*n*	*ACC*	
1	**389**	**0.975**	**389**	**0.975**	**389**	**0.975**	399
2	384	0.962	384	0.962	**389**	**0.975**	
3	**382**	**0.957**	381	0.955	**382**	**0.957**	
4	**387**	**0.970**	381	0.955	**387**	**0.970**	
5	**381**	**0.955**	373	0.935	**381**	**0.955**	
6	**383**	**0.960**	372	0.932	**383**	**0.960**	
7	**379**	**0.950**	372	0.932	**379**	**0.950**	
8	**379**	**0.950**	372	0.932	**379**	**0.950**	
9	**377**	**0.945**	369	0.925	**377**	**0.945**	
10	**377**	**0.945**	369	0.925	**377**	**0.945**	
ecoli -> α = 0.1							
K	A		B		KNN		Total
	n	*ACC*	*n*	*ACC*	*n*	*ACC*	
1	**274**	**0.815**	**274**	**0.815**	**274**	**0.815**	336
2	265	0.789	265	0.789	**274**	**0.815**	
3	285	0.848	**288**	**0.857**	285	0.848	
4	285	0.848	**287**	**0.854**	286	0.851	
5	285	0.848	283	0.842	**286**	**0.851**	
6	**286**	**0.851**	282	0.839	**286**	**0.851**	
7	**287**	**0.854**	286	0.851	**287**	**0.854**	
8	291	0.866	285	0.848	290	0.863	
9	287	0.854	287	0.854	**288**	**0.857**	
10	288	0.857	278	0.827	**289**	**0.860**	

Tabela 1: Resultados para as bases de dados "*ecoli*" e "*compound*" no Primeiro experimento

S4 -> K = 10							
α | α*K	A		B		KNN		Total
	n	*ACC*	*n*	*ACC*	*n*	*ACC*	
0.0 | 0	**2010**	**0.402**	2007	0.401	2010	0.402	**5000**
0.1 | 1	**2010**	**0.402**	2007	0.401			
0.2 | 2	2009	0.402	1999	0.400			
0.3 | 3	**2017**	**0.403**	1991	0.398			
0.4 | 4	**2015**	**0.403**	1985	0.397			
0.5 | 5	**2015**	**0.403**	1992	0.398			
0.6 | 6	**2019**	**0.404**	1992	0.398			
0.7 | 7	**2019**	**0.404**	1992	0.398			
0.8 | 8	1949	0.390	1954	0.391			
0.9 | 9	1939	0.388	1615	0.323			

pathbased -> K = 10							
α | α*K	A		B		KNN		Total
	n	*ACC*	*n*	*ACC*	*n*	*ACC*	
0.0 | 0	**296**	**0.987**	295	0.983	296	0.987	**300**
0.1 | 1	**296**	**0.987**	295	0.983			
0.2 | 2	**296**	**0.987**	295	0.983			
0.3 | 3	**296**	**0.987**	295	0.983			
0.4 | 4	**296**	**0.987**	295	0.983			
0.5 | 5	**296**	**0.987**	295	0.983			
0.6 | 6	295	0.983	294	0.980			
0.7 | 7	295	0.983	294	0.980			
0.8 | 8	290	0.967	280	0.933			
0.9 | 9	285	0.950	222	0.740			

Tabela 2: Resultados para as bases de dados "*S4*" e "*pathbased*" no Segundo Experimento.

Conclusões

Os resultados indicam que o melhor desempenho é geralmente obtido pelo KNN original, ou pelo KNN proposto para um valor de *α* pequeno, em bases de dados bidimensionais e sem sobreposição entre clusters. Nestes casos, escolher os vizinhos mais próximos utilizando densidade afeta o desempenho do KNN proposto de forma neutra ou negativa. Entretanto, melhores resultados são obtidos em bases com dimensões maiores e/ou sobreposição de clusters. Os resultados para os experimentos investigando os impactos de *α* indicam que melhores resultados são geralmente obtidos para menores valores de *α*. A exceção são as bases de dados com sobreposições de clusters, cujos resultados são melhores ao selecionar vizinhos baseados em densidade espacial.

Agradecimento: CNPq pelo apoio financeiro.

Referências Bibliográficas

[1] AHA, D. W.; KIBLER, D. & ALBERT, M.K. (1991). “Instance-based learning algorithms”, Machine Learning, 6(1): 37-66.

[2] DUA, D. & GRAFF, C. (2019). “UCI Machine Learning Repository”, [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

[3] FIX, E. (1985). “Discriminatory analysis: nonparametric discrimination, consistency properties”, Technical Report, USAF School of Aviation Medicine.

[4] FRÄNTI, P. & SIERANOJA, S. (2018). “K-means properties on six clustering benchmark datasets”, Applied Intelligence, 48 (12): 4743-4759.

[5] KOTSIANTIS, S. B.; ZAHARAKIS, I. & PINTELAS, P. (2007). “Supervised machine learning: A review of classification techniques”, Emerging artificial intelligence applications in computer engineering, 160(1): 3-24.

[6] TINÓS, R.; ZHAO, L.; CHICANO, F. & WHITLEY, D. (2018). “NK hybrid genetic algorithm for clustering”, IEEE Transactions on Evolutionary Computation, 22(5): 748-761.