

Uma proposta para valores k Locais para a regra de k -Nearest Neighbor

Nicolás García-Pedrajas, Juan A. Romero del Castillo, and Gonzalo Cerruela-García

2ª Apresentação

Aluno: Lucas de Souza Albuquerque (Isa2)

Outline

- ❖ Introdução
 - Contexto
 - Proposta
- ❖ Trabalhos Relacionados
- ❖ K Local Ótimo para regra KNN
- ❖ Configuração dos Experimentos
- ❖ Resultados
- ❖ Conclusões e Trabalhos Futuros

Introdução - Contexto

- ❖ Um classificador é uma função $f : X \rightarrow Y$ que mapeia uma instância x à uma classe y .
- ❖ KNN (*k-nearest neighbor*) é um método conhecido de classificação:
 - Conjunto de Protótipos (x_i, y_i) representam nosso conhecimento sobre um problema.
 - Classificação de uma instância baseada na dos k vizinhos mais próximos.
- ❖ Porém, o KNN tem um problema.
 - Como escolher o valor ideal para k ?

Introdução - Proposta

- ❖ Valores diferentes de k apresentam diferentes taxas de acerto para um problema.
- ❖ É possível se achar um bom valor para k por meio de Cross-Validation (**CV**)
 - Mas é improvável que esse mesmo valor seja o melhor para todo o espaço do problema.
- ❖ Neste trabalho se é apresentado um método para usar e treinar valores de k locais.
 - Este método é rápido e preciso, melhorando a capacidade de generalização do KNN sem piorar a complexidade.

Introdução - Proposta

- ❖ Cada protótipo recebe um valor k , que representa o valor ótimo de vizinhos à serem usados naquela vizinhança no kNN
 - Então, em vez do formato $(\mathbf{x}_i, \mathbf{y}_i)$, onde \mathbf{x}_i é o protótipo e \mathbf{y}_i sua classe, é se usado protótipo melhorados com formato $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{k}_i)$, onde \mathbf{k}_i é o valor k associado ao protótipo \mathbf{x}_i
- ❖ O método permite escolher um valor k local facilmente e rapidamente.
 - Em tempo de treino, o método possui complexidade linear
 - Em tempo de teste, o método tem a mesma carga de trabalho que o kNN padrão.

Trabalhos Relacionados

- ❖ Existem métodos que falam sobre valores de k não globais
 - Mas nenhum deles mostrou significativa melhoria sobre a abordagem de um valor k global escolhido por CV.
- ❖ Ferrer-Troyano *et al.* propôs o *k-frequent-NN*.
 - O método tenta remover o uso de um k global, mas não se existe um valor local de k para cada instância.
- ❖ Wettschereck e Dietterich desenvolveram um modelo que guarda, para cada instância, uma lista de todos os valores de k que o classificaram corretamente.
 - Não apresentou melhorias sobre o KNN padrão.

...entre outros



K-Local ótimo para a regra KNN

- ❖ A abordagem é baseada em fornecer um valor local k para a vizinhança de cada protótipo.
 - Para uma instância nova de classe desconhecida, o vizinho mais próximo do banco de treinamento é obtido, e o valor k associado à este vizinho é usado para se classificar aquela instância.
- ❖ O processo de treinamento deve então, obter o melhor valor k para cada protótipo.
 - É proposto um algoritmo guloso: para cada protótipo, testa todos os valores de k em um intervalo $[k_{min}, k_{max}]$



K-Local ótimo para a regra KNN

- Para obter o valor local de k associado à um protótipo x_i , só precisamos considerar as instâncias quem tem x_i como vizinho mais próximo.
 - A avaliação dos valores no intervalo $[k_{min}, k_{max}]$, então, é rápida.
- Porém, para alguns protótipos, o número de vizinhos mais próximos pode ser baixo ou até zero.
 - Para se evitar este problema, são considerados partes da 'vizinhança' todas as instâncias que tenham o protótipo como um dos seus n -vizinhos mais próximos. ($n = 3$)
- Para analisar cada valor k , nosso primeiro objetivo é a performance de classificação.
 - Medida como a taxa de acerto para problemas padrões.



K-Local ótimo para a regra KNN

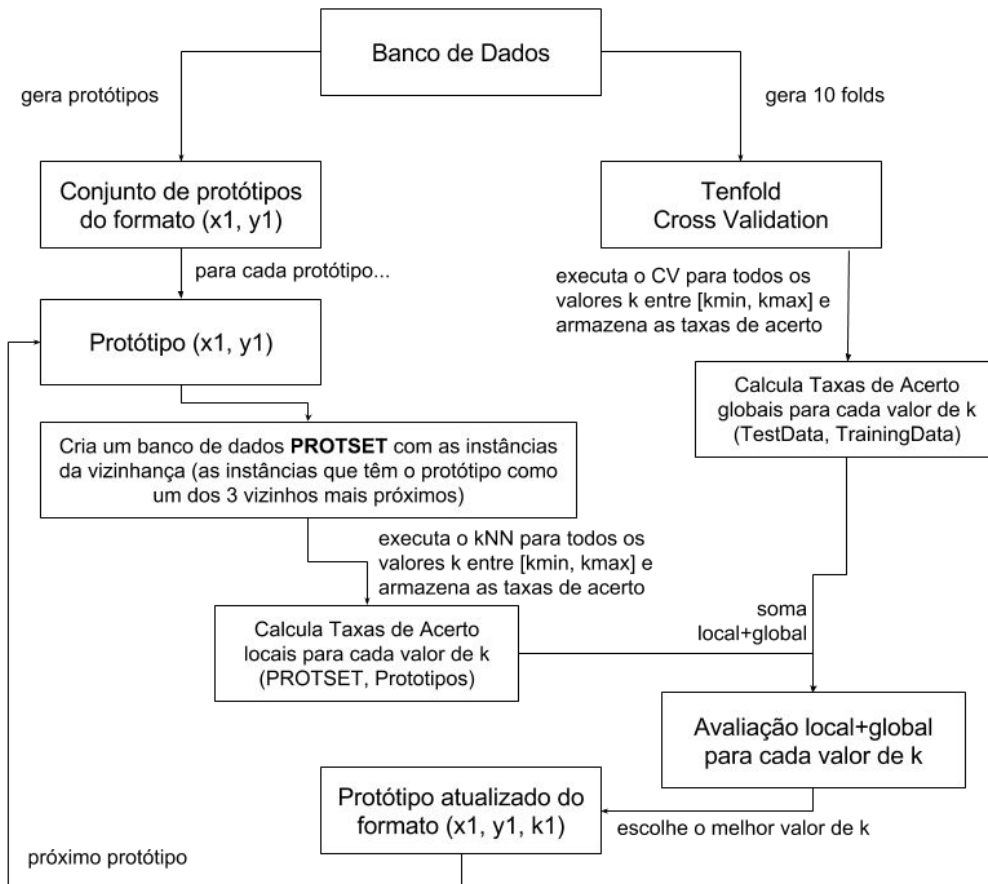
- ❖ Podem se existir vários valores de k com a mesma performance.
- ❖ Neste caso, não se sabe necessariamente qual o melhor valor:
 - Então, se é adicionada a performance global de um valor de k à sua performance local.
 - Esta performance global é obtida por *CV 10-fold*.

A nova avaliação do valor k se dá agora pela soma das taxas acertos globais e locais.

Esta combinação quebra empates e adiciona uma visão global ao sistema, evitando variações grandes de k locais.



K-Local ótimo para a regra KNN




Configuração Experimental

- ❖ Foco no kNN padrão, mas pode se expandir facilmente para outras versões do kNN e outros cálculos de distância.

Nos experimentos:

- Todo k deve estar dentro do intervalo $[1, 10]$
- Valores de k global encontrados por CV 10-Fold
- Protótipos gerados por CNN à partir de uma divisão no banco de dados.

Serão comparados os seguintes algoritmos:

- KNN padrão, representado por **standard KNN**, ou **KNN**
 - KNN com protótipos, representado por **standard PROT**, ou **KNN Prot**
 - Local k com protótipos, representado por **proposed PROT**, ou **Local k Prot**
 - Local k usando o banco de treinamento inteiro, representado por **proposed KNN**, ou **Local k Full**
- 

Configuração Experimental

❖ Datasets:

- Retirados do repositório da **University of California at Irvine (UCI)**

<i>Dataset</i>	<i># Instâncias</i>	<i># Atributos</i>	<i>Ano de Publicação</i>
iris	150	4	1988
wine	178	13	1991
parkinsons	197	23	2008
sonar	208	60	N/A
seeds	210	7	2012
glass	214	10	1987
haberman	306	3	1999
ecoli	336	8	1996
leaf	340	16	2014
ionosphere	351	34	1989

Configuração Experimental

Medidas de Avaliação:

- *Accuracy*: Porcentagem de instâncias corretamente classificadas
- κ de Cohen: Usada para compensar sucessos aleatórios.

$$\kappa = \frac{n \sum_{i=1}^C x_{ii} - \sum_{i=1}^C x_{i \cdot} x_{\cdot i}}{n^2 - \sum_{i=1}^C x_{i \cdot} x_{\cdot i}}$$



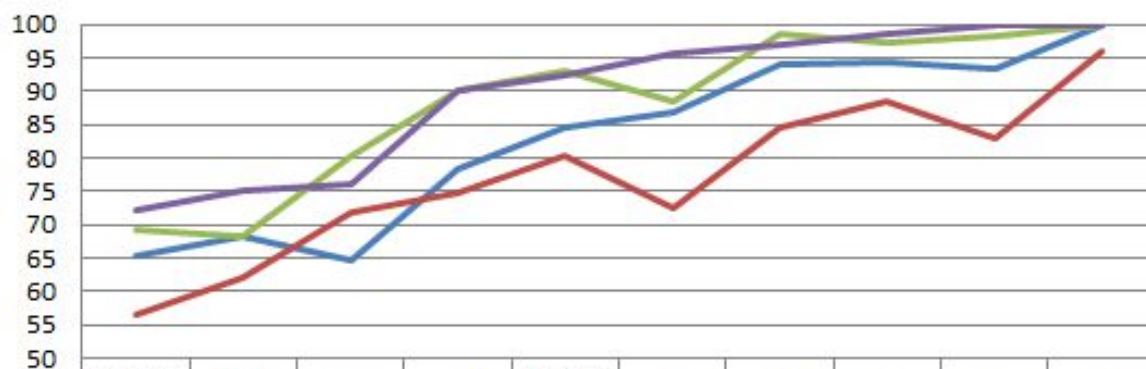
Resultados (Datasets Padrões)

- ❖ Experimentos feitos para comparar o algoritmo proposto com o kNN padrão.
- ❖ Duas colunas para os algoritmos locais: experimento devido à simplicidade da prototipagem
 - Usa o banco de treinamento inteiro como protótipo.
 - Geração de protótipos mais avançado evita isso e deixa o algoritmo ainda mais rápido.
- ❖ Ainda assim, algoritmos mostram melhoria significativa sobre kNN padrão.

<i>Accuracy</i>				
	K-NN	K-NN Prot	Local k Prot	Local k Full
Average	0.829	0.769	0.883	0.897
1 st /2 nd /3 rd /4 th	1/1/7/1	0/0/1/9	5/5/0/0	7/3/0/0
<i>Cohen's κ</i>				
	K-NN	K-NN Prot	Local k Prot	Local k Full
Average	0.701	0.591	0.788	0.809
1 st /2 nd /3 rd /4 th	1/0/8/1	0/0/1/9	6/4/0/0	6/4/0/0

Resultados (Datasets Padrões - Taxa de Acerto)

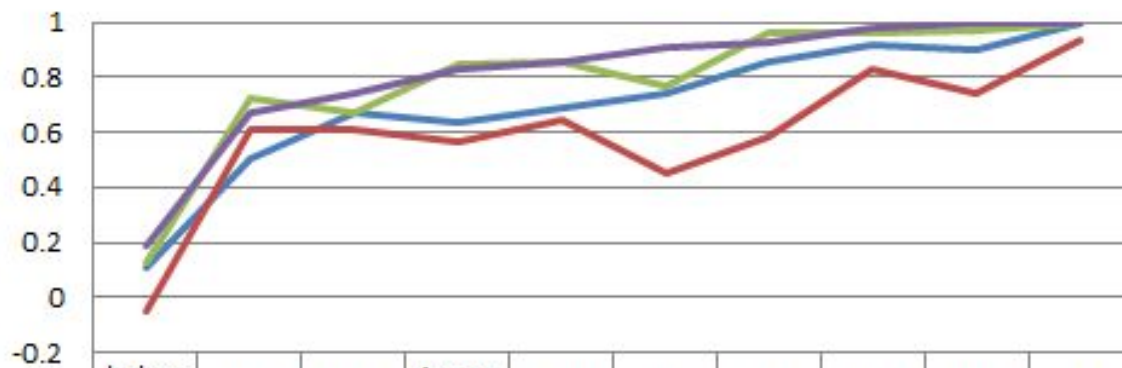
Accuracy



	haberm an	leaf	glass	ecoli	ionosp here	sonar	park	seeds	wine	iris
standard KNN	65.346	68.141	64.788	78.378	84.482	86.956	93.846	94.285	93.22	100
standard PROT	56.435	61.946	71.83	74.774	80.172	72.463	84.615	88.571	83.05	96
proposed PROT	69.306	68.141	80.281	90.09	93.103	88.405	98.461	97.142	98.305	100
proposed KNN	72.277	75.221	76.056	90.09	92.241	95.652	96.923	98.571	100	100

Resultados (Datasets Padrões - κ de Cohen)

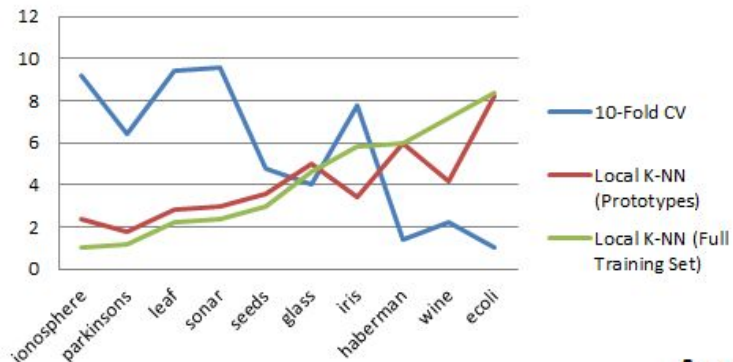
Cohen's Kappa



	haber man	glass	leaf	ionos phere	ecoli	sonar	park	seeds	wine	iris
— standard KNN	0.105	0.506	0.668	0.638	0.689	0.738	0.855	0.914	0.898	1
— standard PROT	-0.047	0.612	0.605	0.564	0.643	0.45	0.58	0.828	0.743	0.939
— proposed PROT	0.128	0.725	0.669	0.845	0.855	0.767	0.963	0.957	0.974	1
— proposed KNN	0.187	0.667	0.742	0.825	0.854	0.912	0.927	0.978	1	1

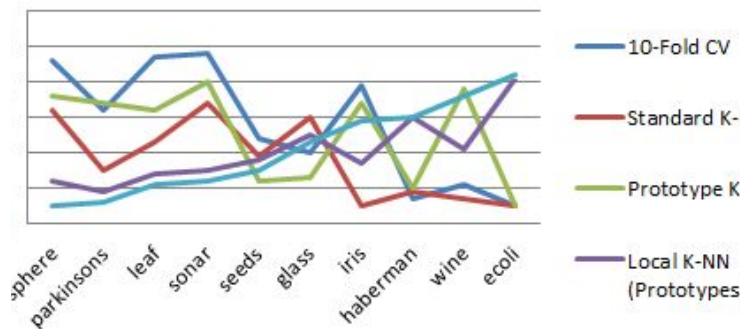
Resultados (Datasets Padrões - k ótimo médio)

Average K



- ❖ Valores do k local tendem à ser o oposto do CV 10-Fold
 - Quando o CV consegue valores altos para k , o algoritmo proposto consegue valores médios mais altos, e vice versa.
 - Uma versão mais extrema do que o visto no artigo original, talvez devido à quantidade de k possíveis ser menor.

Average K



Conclusão

- ❖ Neste trabalho foi-se apresentado um método rápido para introduzir um valor local de k para o classificador KNN
 - Se associa um valor local de k à cada protótipo durante o treinamento baseado na vizinhança do protótipo...
 - ...e instâncias de teste usam o k do vizinho mais próximo.
- ❖ O método foi comparado com a abordagem de achar o k ótimo por CV 10-fold, pelo KNN padrão, e pelo KNN com uso de protótipos, mostrando-se geralmente melhor que estes para todos os bancos de dados avaliados.
 - Em pior dos casos, empata com um dos outros métodos.



Perguntas?

Aluno: Lucas de Souza Albuquerque (Isa2)