Um Estudo Sobre Métricas no Espaço de Similaridade como Preditores de Acurácia para Classificadores KNN

Curso: Otimização

Aluno(a): Eduardo Henrique Basilio de Carvalho

Belo Horizonte, 9 de Julho de 2025

Agenda

Classificador k-Nearest Neighbors (kNN)

- Simples, consolidado e amplamente utilizado;
- Altamente sensível à escolha dos hyperparâmetros:
  - Número de vizinhos k;
  - Raio do Kernel h.

Abordagem Clássica

- A escolha de k e h é feita com base na acurácia do classificador;
- A acurácia é calculada através de validação cruzada;
- O processo é computacionalmente caro, especialmente para grandes conjuntos de dados.

Objetivo do Trabalho

- Investigar se é possível prever a acurácia do classificador kNN com base em métricas do espaço de similaridade;
- Utilizar essas métricas como preditores para determinar os melhores valores de k e h.

Kernel por Função de Base Radial (RBF)

$$K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|^2}{2h^2}\right) \quad (1)$$

- $K(\mathbf{u}, \mathbf{v})$ é o kernel entre os vetores $\mathbf{u}$ e $\mathbf{v}$;
- h é o raio do kernel, que controla a suavidade da função de similaridade.
- O kernel é computado entre cada amostra e seus k vizinhos mais próximos.

Transformação para o Espaço de Similaridade

$$Q_{ik} = \sum_{j=1}^{m} K_{ij} \cdot \mathbb{I}(y_j = c_k) \tag{2}$$

- Q_{ik} é a soma dos pesos das instâncias da classe c_k;
- K_{ij} é o kernel entre as instâncias i e j;
- $\mathbb{I}(y_j = c_k)$ é uma função indicadora que vale 1 se a instância j pertence à classe c_k.

Dissimilaridade

Proposta por Menezes et al. (2019):

$$D_{ij} = |\mathbf{v}i - \mathbf{v}j| \cdot \frac{\mathbf{v}i \cdot \mathbf{v}j}{|\mathbf{v}i||\mathbf{v}j|} \tag{3}$$

Computada entre cada par de centróides. A pontuação é a média das dissimilaridades entre os centróides de cada classe menos o desvio padrão:

$$D = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{C-1}\sum_{c' \neq c} D_{cc'}\right) - \sigma(D_{cc'}) \tag{4}$$

- C é o número de classes;
- $D_{cc'}$ é a dissimilaridade entre os centróides das classes c e c';
- $\sigma(D_{cc'})$ é o desvio padrão das dissimilaridades.

n-Volume da Interseção entre os Fechamentos Convexos

- A pontuação é a média dos volumes das interseções entre os fechamentos convexos de cada par de classes;
- Computada como um problema de otimização linear, resolvido por pontos interiores.

n-Volume dos Fechamentos Convexos

- A pontuação é a média dos volumes dos fechamentos convexos de cada classe menos o desvio padrão;

Separação Espacial

$$\text{Final Score} = ((\mu_{\text{between}} \cdot \mu_{\text{within}}) - (\sigma_{\text{between}} \cdot \sigma_{\text{within}})) \cdot (1 - f_h) \cdot (1 - f_k) \tag{5}$$

- μ_{between} é a média das distâncias entre amostras de classes diferentes;
- μ_{within} é a média das distâncias entre amostras da mesma classe;
- σ_{between} é o desvio padrão das distâncias entre amostras de classes diferentes;
- σ_{within} é o desvio padrão das distâncias entre amostras da mesma classe;
- f_h é um fator de regularização para o raio do kernel;
- f_k é um fator de regularização para o número de vizinhos k.

Silhueta

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{b(i) - a(i)}{\max(a(i), b(i))} \right) \tag{6}$$

- n é o número total de amostras;
- $a(i)$ é a distância média da amostra i para as outras amostras da mesma classe;
- $b(i)$ é a distância média da amostra i para as amostras da classe mais próxima.

Paralelismo Entre o Hyperplano Gerado e o Hyperplano Oposto Unitário

- Mede o paralelismo entre o hiperplano formado pelos centróides das classes e um hiperplano de referência;
- A pontuação é inversamente proporcional à similaridade de cosseno absoluta entre os vetores normais dos dois hiperplanos.

$$\text{Pontuação} = (1 - |\mathbf{n}_{\text{centróide}} \cdot \mathbf{n}_{\text{oposto}}|) \cdot (1 - f_k) \tag{7}$$

- $\mathbf{n}_{\text{centróide}}$ é o vetor normal ao hiperplano dos centróides;
- $\mathbf{n}_{\text{oposto}}$ é o vetor normal ao hiperplano de referência;
- f_k é um fator de regularização para o número de vizinhos k.

Referência

Como referência, um classificador otimizado por validação cruzada k-fold, com $k = 5$ foi utilizado.

Datasets

Tabela: Datasets Utilizados no Estudo

Dataset	Amostras	Atributos	Classes	Fonte
Banknote Authentication	1372	4	2	UCI
Breast Cancer (Wisconsin)	569	30	2	UCI
Car Evaluation	1728	21	4	UCI
Credit Germany (Statlog)	1000	74	2	UCI
Diabetes (Pima)	768	8	2	UCI
Heart Disease (Statlog)	270	13	2	UCI
Ionosphere	351	34	2	UCI

Acurácia Média

Tabela: Acurácia Média do k-NN com Diferentes Métricas

Dataset	Baseline	I	II	III	IV	V	VI
Banknote Auth.	0.942	**0.887**	**0.887**	**0.887**	0.881	**0.887**	**0.887**
Breast Cancer	0.947	0.887	0.899	0.910	**0.913**	0.910	0.887
Car Evaluation	0.850	0.700	0.824	0.701	0.700	**0.850**	0.701
Credit Germany	0.725	0.000	0.000	0.000	0.703	**0.708**	0.000
Diabetes (Pima)	0.751	0.699	0.699	0.723	**0.734**	0.723	0.723
Heart Disease	0.577	0.570	0.570	0.574	**0.587**	0.574	0.580
Ionosphere	0.860	0.717	0.717	**0.780**	**0.780**	0.746	0.717

Tempo de Treinamento

Tabela: Tempo Médio de Treinamento (segundos)

Dataset	Baseline	I	II	III	IV	V	VI
Banknote Auth.	4.599	1.334	1.603	1.659	1.599	1.585	1.568
Breast Cancer	3.841	0.485	0.383	0.409	0.385	0.503	0.516
Car Evaluation	9.923	2.816	3.707	2.663	3.568	3.173	3.236
Credit Germany	14.604	3.980	3.923	3.945	3.872	4.006	4.067
Diabetes (Pima)	4.760	1.158	1.110	0.660	1.027	0.737	0.765
Heart Disease	3.152	1.762	0.618	0.268	0.373	0.291	0.246
Ionosphere	10.724	1.365	1.666	1.425	1.284	1.414	1.511

Tempo de Inferência

Tabela: Tempo Médio de Inferência (segundos)

Dataset	Baseline	I	II	III	IV	V	VI
Banknote Auth.	0.00347	0.0100	0.0124	0.0160	0.0246	0.0108	0.0107
Breast Cancer	0.00257	0.00377	0.00337	0.00360	0.00385	0.00330	0.00390
Car Evaluation	0.0470	0.0366	0.0172	0.0229	0.0420	0.0380	0.0238
Credit Germany	0.298	0.149	0.165	0.153	0.161	0.154	0.165
Diabetes (Pima)	0.00339	0.0110	0.0138	0.00554	0.0120	0.00530	0.00560
Heart Disease	0.00243	0.00254	0.00230	0.00238	0.00240	0.00245	0.00248
Ionosphere	0.286	0.081	0.093	0.070	0.082	0.0611	0.101

Conclusões

- As métricas propostas (I-VI) demonstraram uma redução significativa no tempo de treinamento em comparação com o baseline.
- O impacto na acurácia foi variado. Em alguns datasets, a acurácia foi aproximadamente mantida, mas em outros, houve uma queda de desempenho.
- O tempo de inferência também apresentou resultados mistos, com ganhos de velocidade em alguns casos e perdas em outros.
- Fica evidente um *trade-off* entre velocidade de treinamento e acurácia, indicando que a escolha da métrica depende da prioridade da aplicação.