Classificação com Janelas de Parzen

Savio Lopes Rabelo

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Campus Fortaleza – CE – Brasil

savio.rabelo@ppgcc.ifce.edu.br

Resumo. *Este relatório descreve a utilização de Janelas de Parzen afeta o desempenho de um classificador Bayesiano padrão. A metodologia utilizada para a implementação é constituída por duas fases: treinamento e teste, com cada conjunto sendo composto por 80% e 20% das bases de dados, respectivamente. Foram usadas quatro bases de dados disponíveis online no repositório UCI Machine Learning. Os resultados são bastante satisfatórios, chegando em taxas de acerto em 100% em algumas bases.*

1. Introdução

Esta atividade estuda como a utilização de Janelas de Parzen, da classe de métodos não-paramétricos, afeta o desempenho de um classificador Bayesiano padrão. Métodos como Janelas de Parzen são extremamente sensíveis a escolha do tamanho da janela, portanto, uma busca em grade com validação cruzada de *k-folds* (*grid search with k-fold cross validation*) é realizada para a escolha do tamanho da janela. A regra padrão é definida da seguinte forma:

$$P(W_j|x) = \frac{p(x|W_j)P(W_j)}{p(x)}, \tag{1}$$

onde $p(x|W_j)$ é a função de verossimilhança (*likelihood function*), $P(W_j)$ é a priori e $p(x)$ é a evidência (*evidence*).

Caso se utilize como janela a função gaussiana, temos:

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}h}exp\left\{\frac{-(x_i - x)^2}{2h^2}\right\}, \tag{2}$$

em que h auxilia na definição da janela.

2. Simulações Computacionais

No primeiro momento foi realizada a separação do conjunto de dados em dois subconjuntos: treinamento e teste. Os valores utilizados para os conjuntos equivalem a 80% do conjunto original para a fase de treinamento e 20% do conjunto original para a fase de teste. Logo depois, os dados foram normalizados para eliminação de redundâncias indesejadas e também foram embaralhados. Por ser um problema com mais de 2 classes, foi exigida uma codificação diferente, Um-versus-Todos (do inglês *One-v-All*, OvA).

Além disso, também foi utilizada a busca em grade com validação cruzada *k-fold*. A busca em grade é uma busca com o objetivo de encontrar os melhores parâmetros. Já o método de validação cruzada *k-fold* consiste em dividir o conjunto total de dados em *k* subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é realizado *k* vezes alternando de forma circular o subconjunto de teste. A Figura 1 mostra o esquema realizado pelo *k-fold*. Ao final das *k* iterações calcula-se a acurácia sobre os erros encontrados, obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados.

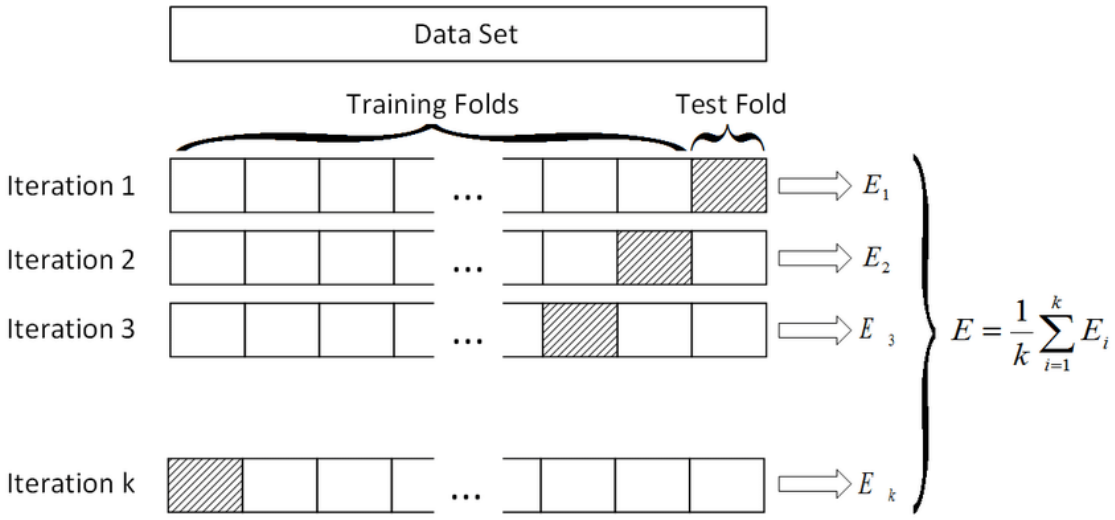

Figura 1. Método k-fold

Para análise comparativas neste estudo, foram usados quatro conjuntos de dados: *Iris Flower Data Set*, *Vertebral Column Data Set*, *Dermatology Data Set* e *Breast Cancer Wisconsin Data Set*; todos disponíveis online no repositório *UCI Machine Learning* [Lichman 2013].

Para realizar os experimentos, foi utilizado um computador com a seguinte configuração: processador Intel(R) Core(TM) i7-6500U a 2.5 GHz com 8 GB de RAM e executando Windowns 10. Além disso, foi utilizado a linguagem de programação MATLAB. Todos os testes foram feitos com 50 realizações em cada base. O tamanho da janela (h) variou de $[0.05\ 0.5]$, com $\Delta h = 0.05$.

Métricas (%)	Bases de Dados					
	Íris	Coluna (3C)	Coluna (2C)	Dermatologia	Câncer	Artificial
Acurácia	95,00	75,29	100,00	95,59	100,00	100,00
Taxa Mínima	86,67	67,74	100,00	90,54	100,00	100,00
Taxa Máxima	100,00	87,10	100,00	100,00	100,00	100,00
Desvio Padrão	3,25	4,25	00,00	2,16	00,00	00,00
Sensibilidade	95,24	70,00	100,00	95,08	100,00	100,00
Especificidade	97,46	87,47	100,00	99,15	100,00	100,00
Precisão	95,20	72,97	100,00	94,98	100,00	100,00
Tempo (s)	22,18	55,36	51,37	94,80	221,45	12,31

Tabela 1. Resultados do classificador Bayesiano padrão com Janelas de Parzen.

Na Figura 2 é apresentada a superfície de decisão com o classificador Bayesiano padrão com Janelas de Parzen.

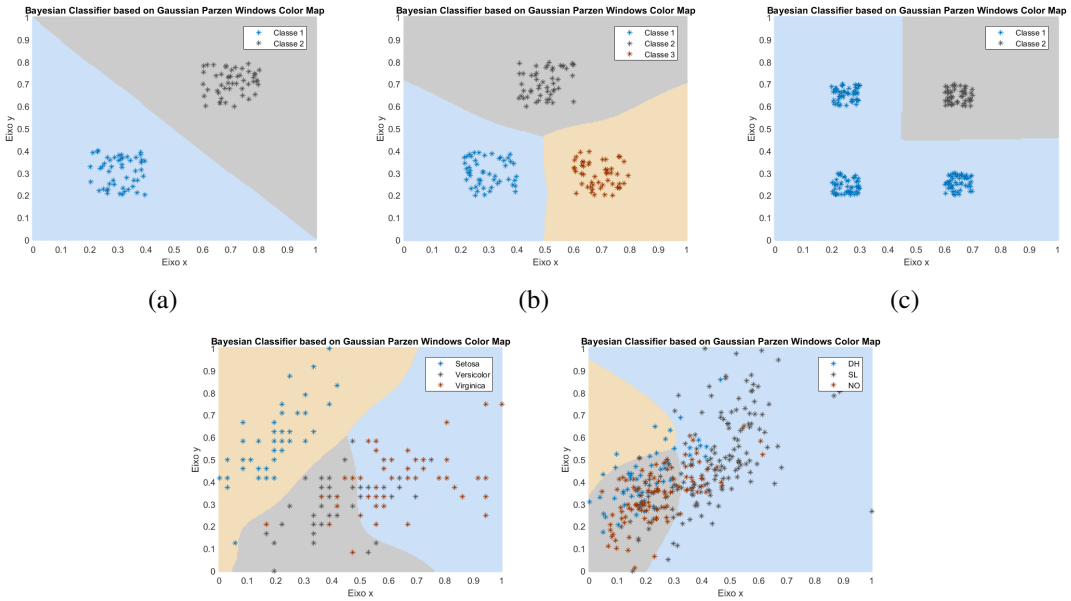

Figura 2. Superfície de decisão do classificador Bayesiano padrão com Janelas de Parzen. (a) Base artificial com 2 classes. (b) Base artificial com 3 classes. (c) Base artificial AND. (d) Íris com dois primeiros atributos. (e) Coluna com dois primeiros atributos.

3. Conclusão

A ideia de adicionar um método não-paramétrico para a estimação da função densidade de probabilidade $p(x)$ é muito bem vista, uma vez que permite um classificador se adaptar a problemas de natureza diversa. Contudo, por não possuir um método de definir o tipo de distribuição a partir do conjunto de dados, precisa-se assumir uma forma de antemão e realizar procedimentos que visam encontrar a melhor parametrização para tal distribuição.

Nesta atividade foi avaliado como um classificador Bayesiano se comporta quando ao invés de utilizar uma forma e parametrização fixas, escolhe-se dar uma flexibilidade maior ao mesmo através de métodos não-paramétricos do tipo janelas de Parzen.

O que foi observado, é que este tipo de solução, embora atrativa, não alterou de forma significante o desempenho de um classificador Bayesiano padrão. Uma das desvantagens é ter que estimar qual o melhor tamanho da janela. Embora a busca em grade com validação cruzada seja um método de avaliação bem definido e robusto, a falta de conhecimento prévio acerca do problema a ser solucionado é um fator que pesa na escolha do intervalo de busca do tamanho. Portanto, o uso de janelas de Parzen para estimação de uma PDF apresenta-se apenas como mais uma técnica útil que tem a seu favor uma flexibilização do classificador em termos de adaptação a problemas diversos.

Referências

Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. Acesso em março de 2019.