Classificador Bayesiano sob Perspectiva Local

Gabriel Baruque, Dereck Torres e Rodrigo Peres
Programa de Pós-Graduação em Engenharia Elétrica
CEFET-RJ
Rio de Janeiro, Brasil
gabriel@baruque.com.br, dereckdemt@gmail.com, rt.peres25@gmail.com

***Resumo*—Desenvolver e aprimorar classificadores é uma tarefa importante em reconhecimento de padrões. Este artigo propõe o desenvolvimento do método Kernel Bayes Local, um classificador inspirado no classificador Bayesiano que utiliza a estimativa de Kernel para as densidades de probabilidade, sendo aplicado em uma abordagem local, obtida através do método não supervisionado K-means. Foram feitos experimentos com 6 bancos de dados, sendo 1 simulado e 5 reais, com o classificador proposto, sua versão global e o KNN. O algoritmo se mostrou bem competitivo quando comparado aos métodos testados e promissor em relação a trabalhos futuros.**

Keywords—Classificador Bayesiano; Classificação Local; Kernel; K-means.

I. INTRODUÇÃO

Aprendizado estatístico [1], [2], engloba métodos de classificação de padrões [3], [4], previsão [1], [5], análise de clusters [2], entre outros. Estes conjuntos de técnicas vêm sendo amplamente aplicadas na literatura científica e no mercado de trabalho, em áreas como medicina [6], análises de redes sociais [7], educação [8], entre outras.

O objetivo em classificação de padrões é, a partir de um conjunto de dados previamente obtido, ajustar um algoritmo a fim de que, se uma nova observação for apresentada, sua classe possa ser estimada. Para isso, ao longo das últimas décadas, diversos métodos foram desenvolvidos e aperfeiçoados, tais como redes neurais [9], support vector machines [10] e random forests [1]. Anterior a estes, o KNN (K-Nearest Neighbors) [2], que classifica uma observação de acordo com a classe mais comum dentre os K vizinhos mais próximos, continua sendo amplamente utilizado devido ao bom desempenho de classificação e facilidade de implementação e interpretação.

Apesar de todo este esforço para desenvolver metodologias eficientes, o classificador Bayesiano [4] é considerado o classificador ótimo, pois possui o menor erro possível. Porém, para calculá-lo é necessário que as densidades de probabilidade dos dados para cada classe sejam conhecidas, o que, em geral, não ocorre. Dessa forma, o uso do classificador Bayesiano, em sua definição, é muito difícil em problemas práticos. Para contornar este problema, pode-se supor que a distribuição por classe obedeça a uma distribuição paramétrica conhecida, como a distribuição normal. Variações desta suposição levam aos discriminantes linear (LDA) e quadrático (QDA) [1]. Outra possibilidade é estimar as densidades através de técnicas não paramétricas, como, por exemplo, kernels [11].

Redes neurais feedforward, LDA e QDA são exemplos de classificadores globais, uma vez que consideram todos os dados disponíveis no conjunto de treinamento para estimar a fronteira de decisão. Em contrapartida, o KNN é um classificador local, já que leva em consideração a informação apenas da vizinhança da observação (os K vizinhos mais próximos) para tomar a decisão de qual classe ela pertence. Em [12], uma proposta de método local para duas classes foi apresentada, onde o k-means, clássico método de análise de clusters, foi usado em uma etapa não supervisionada anterior à classificação, com o objetivo de dividir o espaço da entrada de acordo com a distribuição dos dados. A técnica utilizava os dados localmente cluster a cluster, e nos clusters que continham observações de classes diferentes, uma técnica inspirada no classificador Bayesiano foi aplicada. Uma extensão multiclasse foi publicada em [13].

Classificadores locais podem ser úteis justamente pelo fato de se desconhecer as densidades de probabilidade dos dados por classe. Isto pode levar a um desempenho ruim de métodos globais e, utilizar regiões específicas do espaço de entrada dos dados pode resultar em melhor desempenho.

Neste artigo, apresenta-se um classificador local inspirado no Bayesiano, que pode ser considerado uma extensão da proposta apresentada em [12]. Um procedimento de análise de clusters é realizado, para obter a abordagem local. Em seguida, a análise passa a ser cluster a cluster e a classificação de novas observações dependerá da estrutura local. Se houver uma classe dominante, ela será atribuída a qualquer nova observação alocada ao cluster. Se, ao contrário, houver um equilíbrio maior entre as classes, uma abordagem relacionada ao classificador Bayesiano é usada. Ao contrário de [12], aqui a estimativa é tendo por base a utilização de kernels para as densidades por classe.

A proposta foi testada em 6 bancos, 1 simulado e 5 reais. Comparações foram feitas com o algoritmo local KNN e com a estimativa global por kernel do classificador Bayesiano. Os resultados foram competitivos e o método se mostra promissor.

Pesquisa fomentada pela bolsa de estudos concedida ao aluno Gabriel Baruque.

II. METODOLOGIA

Considere um problema de classificação de padrões com M classes. Seja o conjunto de dados $X = \{x_1, x_2, \dots, x_t | x_i \in \mathbb{R}^n, i = 1, \dots, t; t \in \mathbb{N}\}$ e o vetor Y, cuja observação y_j corresponde a classe da observação x_j, j = 1, ..., t.

O classificador Bayesiano é aquele que apresenta o menor erro de teste entre os classificadores. Considerado o classificador ótimo, é calculado através das probabilidades condicionais de cada classe dada uma observação:

$$Pr(Y = m|X = x) = \frac{\pi_m f_m(x)}{\sum_{l=1}^{M} \pi_l f_l(x)} \quad (1)$$

Onde a probabilidade de uma observação x pertencer a uma classe m depende da estimativa da probabilidade a priori da classe m, π_m, da densidade de probabilidade de x provinda da m-ésima classe, $f_m(x)$, e do somatório das mesmas estimativas provindas de cada classe. A classe de maior probabilidade em (1) é atribuída a X = x.

Apesar de ser o classificador ótimo, o fato de necessitar das densidades de probabilidade dos dados por classe se torna um problema, pois, em geral, são desconhecidas. O método mais comum para se utilizar o classificador Bayesiano é supor uma distribuição paramétrica específica, como a normal, o que leva às definições de discriminantes linear e quadrático, como foi mencionado na introdução. Entretanto, se as distribuições originais não forem normais, os resultados podem ser insatisfatórios.

Uma possibilidade para resolver este problema é utilizar métodos que estimam a distribuição de probabilidade dos dados. Neste trabalho foi utilizado o método de estimativa por Kernel (Janelas de Parzen). Tal método atribui um peso maior a observações próximas de um dado x, decaindo de forma gradual. Dessa forma, a média das distribuições de cada dado pode se tornar muito próxima da distribuição original. Esse método é descrito segundo a seguinte fórmula:

$$\hat{f}_m(x) = \frac{1}{t} \sum_{i=1}^{t} G_{\Sigma_m}(x - x_i) \quad (2)$$

onde G é a função da distribuição normal multivariada, t é o total dos dados e Σ_m é a matriz de covariância da classe m (sendo utilizado nos cálculos, uma matriz diagonal com a variância amostral de cada atributo). Além disso, x é uma observação para a qual o valor na densidade deve ser estimado e x_i são as observações disponíveis.

A estimativa com kernel $\hat{f}_m(x)$ foi utilizada no lugar da função paramétrica $f(x)$ de (1), nos fornecendo a equação utilizada para o classificador Bayesiano:

$$Pr(Y = m|X = x) = \frac{\hat{\pi}_m \hat{f}_m(x)}{\sum_{l=1}^{K} \hat{\pi}_l \hat{f}_l(x)} \quad (3)$$

onde $\hat{f}$ é a estimativa calculada pela equação (2) e $\hat{\pi}$, a estimativa da probabilidade a priori, calculada a partir da frequência da classe. A classe de maior probabilidade em (3) é atribuída a X = x. Esta abordagem é chamada, neste artigo, de Kernel Bayes, já que se trata da estimativa do classificador Bayesiano por kernel.

A fim de implementar a abordagem local, o algoritmo de clusterização k-means será usado. O objetivo deste algoritmo é agrupar as observações em r clusters $\{C_1, C_2, \dots, C_r\}$, onde $C_1 \cup C_2 \cup \dots \cup C_r$ = X e $C_i \cap C_j = \emptyset$, $\forall i, j = 1, \dots, r; i \neq j$. Após a clusterização, o processo de classificação será realizado localmente e com critérios diferentes dentro de cada cluster.

Por ser um processo de clusterização não supervisionado, espera-se que o k-means agrupe dados similares entre si e diferentes dos dados de outros clusters. O objetivo é analisar cada cluster separadamente, resultando, daí, o classificador local.

Para se implementar esse método de clusterização, é necessário fornecer a quantidade de grupos. Esse é um problema bem conhecido na literatura, e foi escolhido o método do cotovelo. Tal método se baseia na soma dos erros quadráticos entre os dados de cada cluster.

Cada cluster obtido foi classificado de acordo com os seguintes parâmetros: homogêneo – caso a classe majoritária contida nele representasse 95% ou mais das observações e outras classes não ultrapassassem 20 observações; heterogêneo – caso a condição anterior não fosse observada. Considera-se então que observações de teste atribuídas a um cluster homogêneo pertencem à mesma classe majoritária do cluster.

Caso sejam atribuídas a um cluster heterogêneo, o classificador Kernel Bayes (equação 3) será usado apenas para as observações desse cluster, o que significa que as estimativas tanto das probabilidades a priori, quanto das densidades serão realizadas através destas observações.

Este método será chamado de Kernel Bayes Local e seu pseudocódigo encontra-se na Fig. 1.

Observe que o Kernel Bayes Local utiliza a mesma metodologia do Kernel Bayes, com a diferença da abordagem local. Pode-se dizer que o Kernel Bayes é um método global, uma vez que utiliza todas as observações disponíveis para o cálculo das estimativas.

- Encontrar a quantidade de clusters
- Clusterizar os dados de treino
- Calcular as probabilidades *a priori* de cada classe para cada cluster
- Definir a homogeneidade ou heterogeneidade do cluster
- Calcular a distância dos dados de teste aos centros dos clusters
- Atribuir os dados de teste ao cluster com centro mais próximo
- Para cada dado de teste
  - Se o cluster cujo dado foi atribuído é homogêneo, então:
    - O dado é classificado com a classe do majoritária do cluster
  - Senão:
    - Estimar a probabilidade através do Kernel Bayes (equação 3)
    - Atribuir a classe com maior probabilidade ao dado verificado
- Comparar as classes reais e estimadas e calcular o erro

Fig. 1. Pseudocódigo do método Kernel Bayes Local

III. Resultados e discussões

São apresentados resultados referentes a um banco simulado e cinco bancos de dados reais, extraídos do repositório da UCI (University of California, Irvine) [14], comparando o desempenho de classificação das observações levando em conta 3 métodos diferentes: Kernel Bayes, Kernel Bayes Local e KNN (K-Nearest Neighbors).

O algoritmo KNN foi escolhido por ser um classificador local, muito bem fundamentado, de fácil implementação, conceitualmente simples e com desempenho eficiente em muitos bancos de dados.

O classificador foi testado com 3, 5, 7 e 9 vizinhos nos dados de treino, para os 6 bancos, individualmente. A configuração que apresentou o melhor desempenho no treinamento foi utilizada também para os dados de teste do respectivo banco.

O banco de dados simulado foi gerado através de 5 normais bidimensionais, centradas nos pontos $(15,15)$, $(15,-15)$, $(0,0)$, $(-15,5)$, $(-15,-5)$ com suas matrizes de covariância $\begin{bmatrix}1 & 2\\ 0 & 18\end{bmatrix}$, $\begin{bmatrix}1 & 2\\ 0 & 18\end{bmatrix}$, $\begin{bmatrix}10 & 6\\ 1 & 1\end{bmatrix}$, $\begin{bmatrix}1 & 2\\ 0 & 18\end{bmatrix}$ e $\begin{bmatrix}1 & 2\\ 0 & 18\end{bmatrix}$ respectivamente. Cada classe contém 100 observações e o conjunto de teste foi gerado com a mesma proporção utilizada nos bancos reais ($\frac{2}{3}$ dos dados foram utilizados para treinamento). Pode-se observar, na Fig. 2, que há uma sobreposição entre duas das cinco classes. Na Tabela I, é mostrado o desempenho dos 3 métodos para o conjunto de teste. As estimativas das matrizes de covariância dos dados utilizadas nos cálculos, seguiram o mesmo princípio mencionado anteriormente.

Observe que os resultados dos 3 métodos foram competitivos. Neste caso, não houve diferença entre a abordagem local e a global.

Os bancos de dados reais utilizados foram: "Data Banknote Authentication" (A), "Wireless Indoor Localization" (B), "Breast Cancer Wisconsin" (C), "Seismic-bumps" (D) e "Yeast" (E).

O banco A apresenta classificação dicotômica, assim como os bancos C e D. Já o banco B apresenta um problema com 4 classes. O banco de dados "Yeast", que originalmente contava com 10 classes, possui uma distribuição muito desbalanceada, onde algumas classes possuíam pouquíssimos elementos.

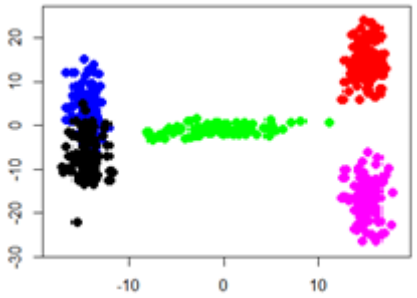

Fig. 2. Distribuição dos dados simulados

TABELA I. Resultado das classificações para o banco simulado

	Acerto (%)
Kernel Bayes	91,62
KNN	**92,22**
Kernel Bayes Local	91,62

Neste caso, a classe com mais elementos foi considerada classe 1 e todas as outras, classe 2. Também, 2 atributos foram retirados, pois apresentavam variância aproximadamente zero, o que impacta no método proposto. Além disso, é de se esperar que atributos que praticamente não variam não irão influenciar na classificação.

A divisão dos dados foi feita seguindo uma proporção onde $\frac{2}{3}$ deles foram destinados a treino e $\frac{1}{3}$ destinado a teste.

Algumas características dos bancos são mostradas na Tabela II.

Na Tabela III tem-se o desempenho dos 3 métodos para os respectivos conjuntos de teste. O número de clusters utilizados no Kernel Bayes Local foi obtido através de testes com os dados dentro da amostra, onde variou-se a quantidade de clusters de 1 a 10, e após, 15, 20, 25 e 30. O melhor resultado para estes dados foi utilizado então nos dados de teste. A única exceção foram os bancos C e D, cujo valor máximo foi de 9 clusters, em virtude da pouca quantidade de observações dentro da amostra.

Do banco A ao E, o número de clusters utilizados foi respectivamente: 25, 30, 6, 9 e 9. Em caso de empate no desempenho, optou-se por escolher a maior quantidade, uma vez que o método se torna mais simples e rápido.

Com os resultados obtidos nos bancos de dados simulado e reais, nota-se que o desempenho dos 3 métodos é próximo. Comparando o Kernel Bayes Local com o Kernel Bayes, houve ganho no banco A. Pode-se considerar praticamente empate em B, C, D e E entre estes dois métodos. Já com o KNN, houve praticamente um empate em A e B, e o método proposto perdeu por menos de 3% nos outros. É verdade que nos resultados próximos a empate, o método proposto sempre ficou

TABELA II. Características dos Bancos Reais

	#Dados	Atributos	Classes	Clusters (método do cotovelo)
A	1372	4	2	5
B	2000	7	4	4
C	683	9	2	4
D	540	18	2	4
E	1484	6	2	4

TABELA III. PERCENTUAL DE ACERTO DAS CLASSIFICAÇÕES

	Kernel Bayes	KNN	Kernel Bayes Local
A	91,48	**100**	99,56
B	98,50	**98,65**	98,20
C	94,74	**96,93**	94,74
D	92,78	**93,89**	91,67
E	69,10	**70,51**	68,69

um pouco abaixo dos demais.

Os resultados aqui apresentados são preliminares. Embora haja um equilíbrio muito grande entre os 3 métodos, alguns pontos podem ser destacados. Na execução do algoritmo, foi observado que o KNN apresentou o processamento mais rápido, porém o Kernel Bayes Local obteve um tempo menor que o Kernel Bayes. Isso ocorre porque o método global estima as densidades para todas as observações de teste, enquanto o Kernel Bayes Local, apenas para as observações alocadas a clusters heterogêneos. Isto indica uma vantagem do método proposto sobre o Kernel Bayes e inclusive sobre o KNN, uma vez que todas as distâncias necessitam ser calculadas para cada observação de teste neste método.

IV. CONCLUSÃO E TRABALHOS FUTUROS

Neste artigo foi proposto um algoritmo denominado Kernel Bayes Local, inspirado no classificador Bayesiano, que utiliza o algoritmo k-means para uma abordagem local. As estimativas das densidades de probabilidade são feitas através do uso de kernels e as probabilidades a priori são estimadas por frequência. O procedimento local é promissor, se mostrando competitivo com o KNN, clássico algoritmo local que costuma ter desempenho muito bom, e com o algoritmo global Kernel Bayes. Ainda, o método proposto é uma maneira mais rápida de se implementar o classificador Bayesiano, poupando tempo de processamento para bancos maiores, com relação ao Kernel Bayes.

Como proposta para trabalhos futuros, é interessante apresentar pelo menos um banco de dados simulado onde o método se destaque em relação aos demais, podendo-se considerar a princípio, bancos desfavoráveis à classificação, seja por dificuldade ou rótulos incorretos. A intuição visual, que embora em muitos casos seja irrelevante ao se aumentar a dimensão, pode ser uma excelente ilustração do método. Além disso, realizar testes com mais bancos de dados e mais comparações, além da extensão do método para o problema de seleção de observações, são propostas viáveis.

AGRADECIMENTOS

Os autores agradecem à professora Caroline Ponce pela ajuda com o software RStudio.

REFERÊNCIAS

[1] T. Hastie, R. Tibshirani e J. H. Friedman, "The Elements of Statistical Learning". Springer, 2ª edição, 2001.

[2] G. James, D. Witten, T. Hastie, R. Tibshirani. "An introduction to Statistical Learning", Springer, 2013.

[3] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2011.

[4] R. O. Duda, P. E. Hart e G. Stork. Pattern Classification, Wiley, 2ª edição, 2001.

[5] D. C. Montgomery e G. C. Runger, Applied Statistics and Probability for Engineers, Wiley, 6ª edição, 2013.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, et al. "A survey on deep learning in medical image analysis"in Medical Image Analysis, 42, pp. 60–88, 2017.

[7] Z. Sun, L. Han, W. Huang, et al. "Recommender systems based on social networks" in Journal of Systems and Software, 99, pp. 109-119, Janeiro 2015.

[8] C. Romero e S. Ventura. "Educational data mining: A survey from 1995 to 2005" in Expert Systems with Applications, 33, pp. 135–146, Julho 2007.

[9] S. S. Haykin, "Neural Networks: A Comprehensive Foundation", Prentice Hall, 1998.

[10] V. N. Vapnik, "Statistical Learning Theory", New York, Wiley, 1998.

[11] B. W. Silverman, "Density Estimation for Statistics and Data Analysis" in Monographs on Statistics and Applied Probability 26, Chapman & Hall/CRC, 1986.

[12] R. T. Peres, C. E. Pedreira, "A new local-global approach for classification", Neural Networks, 23, 2010.

[13] R. T. Peres e C. E. Pedreira, "Novas análises e experimentos em modelo local-global para classificação" in XI Congresso Brasileiro de Inteligência Computacional, 2013.

[14] Repositório UCI: https://archive.ics.uci.edu/ml/index.php (último acesso em 01/05/2019).