Seleção de Observações Aplicada à Classificação

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

***Abstract*—Atualmente, técnicas na área de aprendizado estatístico vêm sendo amplamente utilizadas tanto em pesquisas científicas quanto em aplicações no mercado e indústria. Em função da grande quantidade de dados gerada com a informatização de sistemas, aplicações em Big Data demandam desenvolver e aprimorar metodologias capazes de extrair informações úteis a partir de grandes conjuntos de dados. Este artigo tem como proposta o desenvolvimento de um método de seleção de observações do conjunto de treinamento. O objetivo é verificar se é possível manter o desempenho de classificação utilizando quantidades cada vez menores de dados. O algoritmo utiliza uma estimativa do classificador Bayesiano para determinar as observações que serão selecionadas. Este procedimento foi aplicado ao algoritmo de vizinhos mais próximos (KNN) e aplicações foram realizadas em três bancos de dados, com bom desempenho do método. Os resultados se mostraram positivos em manter e até melhorar a classificação utilizando as observações selecionadas.**

Keywords— Classificador Bayesiano; Kernel; Seleção de Observações; KNN.

I. INTRODUÇÃO

Aprendizado estatístico [1], [2] contempla métodos de reconhecimento de padrões [3], [4], previsão, [1], [5] clusterização [2], dentre outros. Diversas aplicações vêm sendo apresentadas em áreas como medicina [6], análises de redes sociais [7], entre outras.

Em seleção de observações, o objetivo é selecionar os dados mais importantes do conjunto de treinamento para uma determinada tarefa ou análise. Essa área vem sendo amplamente estudada, uma vez que suas aplicações podem ser vistas em sistemas de classificação de imagens, utilizando Active Learning [8], [9], problemas de otimização dinâmica, com Query Based Learning [10] e ainda pesquisas para se desenvolver melhores classificadores, como por exemplo o DROP3 [11].

Apesar de todo o esforço para se desenvolverem métodos eficientes, o classificador bayesiano [4] é ainda considerado o classificador ótimo em termos de menor erro de classificação. Porém para utiliza-lo é necessário conhecer a densidade de probabilidade dos dados para cada classe, e em problemas reais, isso geralmente não acontece. Para resolver este problema, uma das abordagens é considerar uma distribuição paramétrica conhecida, como a normal, para a distribuição dos dados. Esta consideração leva aos discriminantes linear (LDA) e quadrático (QDA) [1] através de variações. Estimar a densidade através de técnicas não paramétricas, como kernels [12] é uma outra possibilidade. A estimativa do classificador bayesiano com kernels é utilizada neste artigo com a proposta de selecionar as observações mais importantes no banco de dados.

Um método de seleção de observações é proposto neste artigo. Para isso, a estimativa de Bayes com Kernel normal é utilizada a fim de se estimar a probabilidade de cada observação do conjunto de treinamento em relação a sua classe. Espera-se que observações típicas de uma classe tenham uma estimativa consideravelmente maior que observações de outras classes, de fronteira, ou ruidosas. Estes dados de treino foram divididos em categorias baseadas nestas estimativas, e grupos de cada combinação de categorias foram criados. Tais grupos foram utilizados no classificador KNN para que se pudesse avaliar seu desempenho ao classificar os dados de validação. Por fim, o grupo de melhor desempenho foi utilizado para classificar os dados de teste. Este procedimento foi realizado em 3 bancos de dados, extraídos do repositório UCI [13] e comparações com seleções aleatórias de dados foram feitas.

Os objetivos principais deste trabalho são: verificar se é possível manter o desempenho de classificação com subconjuntos cada vez menores do conjunto de treinamento e identificar as observações mais importantes de cada classe para o desempenho de classificação.

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 apresenta o menor erro de teste entre os classificadores. É considerado o classificador ótimo, e calculado através das probabilidades condicionais de cada classe dada uma observação:

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

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

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 a partir 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 considerado ótimo, o fato de necessitar das densidades de probabilidade dos dados por classe se torna um impedimento, pois geralmente, 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 forma de resolver este problema é a utilização de métodos que estimam a distribuição de probabilidade dos dados. Neste trabalho foi utilizado o método de estimativa por Kernel (Janelas de Parzen). Este 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:

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

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:

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

em que $\hat{f}$ 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.

Essa estimativa foi calculada para cada observação de treino. Quanto mais próximo de 1 for o resultado da estimativa, mais confiança haverá de que a observação seja realmente daquela classe, ou seja, um dado da classe 1 que esteja inserido em uma área majoritariamente de classe 2 terá uma estimativa baixa. As observações das classes são categorizadas, separadamente, de acordo com o resultado dessa estimativa da seguinte forma:

- Grupo 1: dados com probabilidade estimada menor ou igual a 0.2
- Grupo 2: dados com probabilidade estimada entre 0.2 e 0.4
- Grupo 3: dados com probabilidade estimada entre 0.4 e 0.6
- Grupo 4: dados com probabilidade estimada entre 0.6 e 0.8
- Grupo 5: dados com probabilidade estimada maior que 0.8

Foram utilizadas exclusivamente as observações de cada combinação desses grupos para classificar os dados de validação com o KNN, utilizando 3, 5, 7 e 9 vizinhos. O número de vizinhos com melhor taxa de acerto foi atribuído para cada combinação de grupos. Essa taxa de acerto foi utilizada para gerar o mapa de cores, demonstrando os resultados citados para cada grupo. O mapa é formado por 25 regiões, cada uma representando uma combinação dos grupos citados e sua respectiva cor, representando a taxa de acerto da classificação dos dados de validação pelo KNN.

Com o mapa de cor, foi escolhida a combinação que obteve maior êxito na classificação dos dados de validação, e foi usada para classificar os dados de teste, novamente através do KNN com o mesmo número de vizinhos prévio. Diminuiu-se a quantidade de dados de treino (utilizados com o KNN) gradativamente, sendo escolhidos aleatoriamente de dentro da combinação. Cada classificação foi realizada 50 vezes e resultados médios são apresentados. Para comparação, a classificação com KNN foi feita utilizando-se a mesma quantidade de dados, sendo estes escolhidos aleatoriamente de todo o conjunto de treinamento.

Três bancos de dados foram selecionados a partir do repositório da UCI. São eles: "Waveform Version 2" (A), "Anuran Calls" (B) e "Electrical Grid Stability" (C), e suas características podem ser visualizadas na Tabela I. Modificações foram feitas para que os bancos possuíssem 2 classes, uma sendo representada pelos dados da classe originalmente mais populosa, e a outra com os demais dados. Os dados foram divididos em 3 partes iguais entre treinamento, validação e teste.

TABELA I. CARACTERÍSTICAS DOS BANCOS DE DADOS

	#Dados	Atributos
A	5000	40
B	7195	22
C	10000	12

III. RESULTADOS E DISCUSSÕES

O método proposto foi aplicado nos três bancos de dados descritos. Em seguida, foi gerado um erro de 20% dos dados em cada banco, isto é, os rótulos de saída de 20% das observações de cada banco foram deliberadamente trocados. Após essa troca, o processo de seleção de observações foi realizado novamente e os resultados são apresentados.

A. *Waverform version 2 (A)*

Os resultados da classificação dos dados de validação sugerem que as observações que obtiveram um valor acima de 0,8 no classificador Kernel Bayes são aquelas que influenciam o KNN para uma melhor classificação. Uma representação visual pode ser vista na Fig. 1 através do Mapa de Cor. Os resultados na Tabela II mostram que mesmo utilizando-se uma quantidade gradativamente menor de dados, o KNN obteve desempenho semelhante, ligeiramente menor, utilizando dados de treinamento aleatórios, ao classificar os dados de teste.

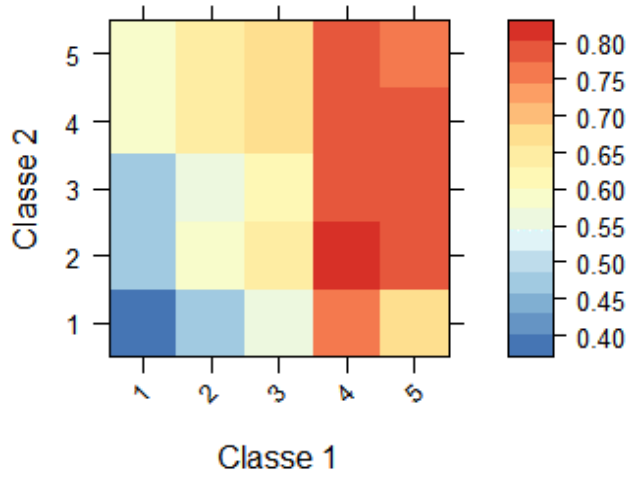

Fig. 1. Mapa de cor do Banco A (com troca de rótulo)

TABELA II. DESEMPENHO DE CLASSIFICAÇÃO NO BANCO A

	Método proposto		Seleção aleatória	
# obs	Média de desempenho	Desvio padrão	Média de desempenho	Desvio padrão
1000	**0.86**	0.00	0.83	0.01
800	**0.86**	0.00	0.83	0.01
600	**0.86**	0.01	0.83	0.01
400	**0.85**	0.01	0.82	0.01
200	**0.84**	0.01	0.81	0.01
180	**0.84**	0.01	0.81	0.02
160	**0.83**	0.01	0.81	0.02
140	**0.83**	0.01	0.80	0.02
120	**0.83**	0.02	0.80	0.02
100	**0.82**	0.02	0.80	0.02
80	**0.81**	0.03	0.79	0.02
60	**0.80**	0.03	0.78	0.02
40	**0.79**	0.04	0.76	0.03
20	0.72	0.05	0.72	0.04
15	**0.70**	0.04	0.69	0.05

Com a inserção da troca de rótulo neste dataset, mudanças ocorreram tanto nos dados selecionados (observar Fig. 2), quanto nos resultados da classificação dos dados de teste. Os dados selecionados desta vez foram aqueles com a estimativa Kernel Bayes entre 0,6 e 0,8 para classe 1, e 0,2 e 0,4 para classe 2. Neste cenário, o método proposto obteve um desempenho melhor, com uma maior taxa de acerto na classificação, apresentada na Tabela III.

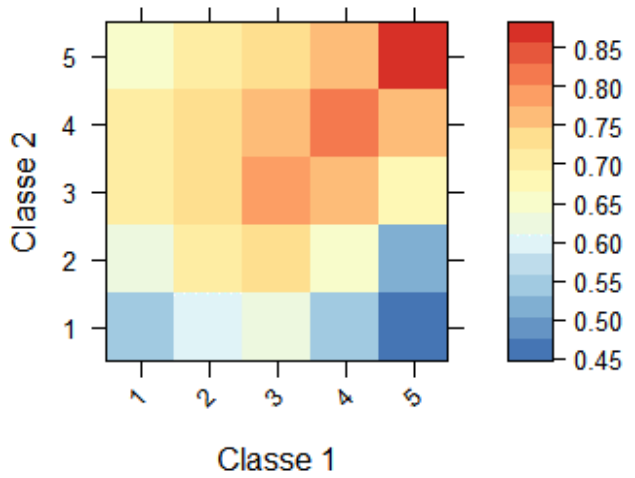

Fig. 2. Mapa de cor do Banco A

B. *Anuran Calls (B)*

Neste banco de dados, para o experimento sem troca de rótulo, observa-se na Fig. 3 que os grupos que obtiveram melhores resultados para a classificação dos dados de validação foram aqueles com estimativa maior que 0,8 em Kernel Bayes para ambas as classes. As comparações realizadas com o

TABELA III. DESEMPENHO DE CLASSIFICAÇÃO NO BANCO A COM TROCA DE RÓTULO

	Método proposto		Seleção aleatória	
# obs	Média de desempenho	Desvio padrão	Média de desempenho	Desvio padrão
240	**0.78**	0.01	0.72	0.05
220	**0.78**	0.01	0.72	0.05
200	**0.78**	0.01	0.71	0.05
180	**0.78**	0.01	0.72	0.05
160	**0.77**	0.02	0.72	0.05
140	**0.76**	0.02	0.71	0.05
120	**0.76**	0.03	0.72	0.05
100	**0.74**	0.03	0.68	0.07
80	**0.73**	0.03	0.70	0.06
60	**0.72**	0.05	0.67	0.09
40	**0.69**	0.07	0.66	0.10
20	0.57	0.13	**0.63**	0.11
15	**0.56**	0.13	0.54	0.13

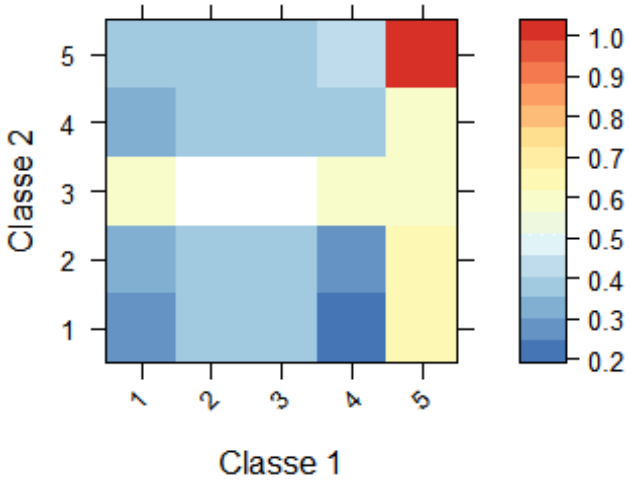

Fig. 3. Mapa de cor do Banco B

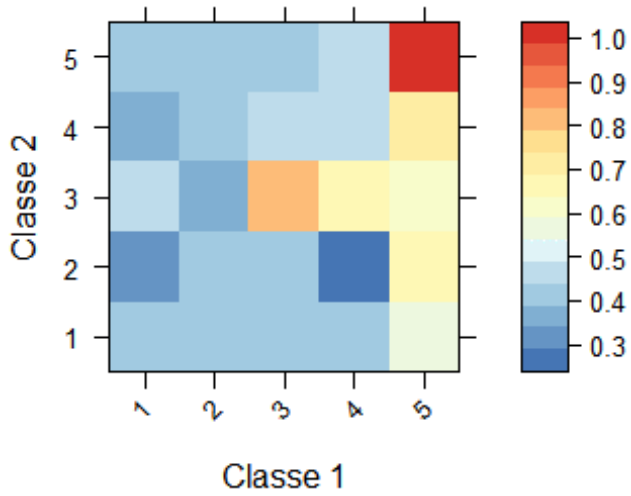

Fig. 4. Mapa de cor do Banco B (com troca de rótulo)

classificador KNN indicam que neste caso, a seleção de observações do método proposto obteve uma taxa de acerto da classificação dos dados de teste, em geral, menor (Tabela IV).

Na Fig. 4 podem ser observadas as mudanças no mapa de cor quando a troca de rótulos é inserida no banco. Nesse caso, os dados foram selecionados na mesma região. O método proposto obteve resultados melhores na classificação das observações de teste, expostos na Tabela V.

TABELA IV. DESEMPENHO DE CLASSIFICAÇÃO NO BANCO B

	Método proposto		Seleção aleatória	
# obs	média de desempenho	desvio padrão	média de desempenho	desvio padrão
2200	0.98	0.00	**0.99**	0.00
2000	0.98	0.00	**0.99**	0.00
1800	0.98	0.00	**0.99**	0.00
1600	0.98	0.00	**0.99**	0.00
1400	0.98	0.00	**0.99**	0.00
1200	0.98	0.00	**0.99**	0.00
1000	0.98	0.00	**0.98**	0.00
800	0.98	0.00	**0.98**	0.00
600	0.98	0.00	**0.98**	0.00
400	0.97	0.00	0.97	0.00
200	0.96	0.01	0.96	0.01
180	0.96	0.01	0.96	0.01
160	0.96	0.01	0.96	0.01
140	0.95	0.01	0.95	0.01
120	0.95	0.01	0.95	0.01
100	0.94	0.01	**0.95**	0.01
80	**0.94**	0.01	0.93	0.02
60	**0.93**	0.01	0.92	0.02
40	0.90	0.03	0.90	0.03
20	0.86	0.04	0.86	0.05
15	0.82	0.07	**0.83**	0.06

TABELA V. DESEMPENHO DE CLASSIFICAÇÃO NO BANCO B COM TROCA DE RÓTULO

	Método proposto		Seleção aleatória	
# obs	Média de desempenho	Desvio padrão	Média de desempenho	Desvio padrão
1800	**0.98**	0.00	0.85	0.01
1600	**0.98**	0.00	0.85	0.01
1400	**0.98**	0.00	0.84	0.01
1200	**0.98**	0.00	0.84	0.01
1000	**0.98**	0.00	0.84	0.02
800	**0.98**	0.00	0.83	0.02
600	**0.98**	0.00	0.83	0.02
400	**0.97**	0.00	0.84	0.03
200	**0.96**	0.01	0.81	0.04
180	**0.96**	0.01	0.81	0.05
160	**0.96**	0.01	0.81	0.05
140	**0.95**	0.01	0.82	0.04
120	**0.95**	0.01	0.81	0.05
100	**0.94**	0.02	0.80	0.05
80	**0.93**	0.02	0.79	0.07
60	**0.91**	0.03	0.77	0.06
40	**0.90**	0.03	0.78	0.07
20	**0.87**	0.03	0.72	0.14
15	**0.85**	0.05	0.69	0.15

C. Electrical grid Stability (C)

Para este banco de dados, nota-se que as observações selecionadas se encontram em uma região diferente dos demais bancos, como pode ser visto na Fig. 5. O grupo de dados que se saiu melhor na classificação dos dados de validação, foi aquele com estimativas de Kernel Bayes entre 0,4 e 0,6 para a classe 1, e acima de 0,8 para a classe 2.. Os resultados da classificação dos dados de teste são vistos na Tabela VI, que demonstra pequeno ganho na taxa de acerto da classificação ao se utilizar o método proposto, independentemente da quantidade de observações utilizadas para se classificar os dados.

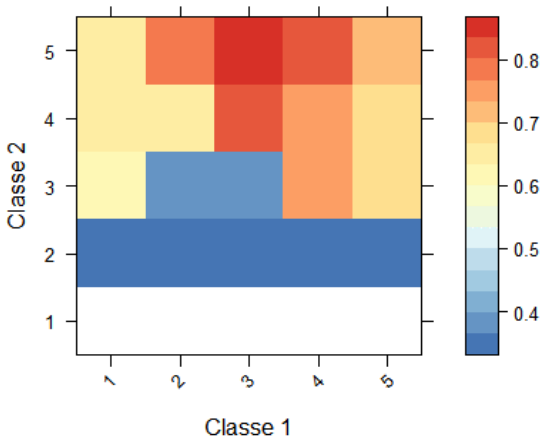

Fig. 5. Mapa de cor do Banco C

Quando inserida a troca de rótulo no banco de dados, o novo grupo de dados escolhido foi aquele com Kernel Bayes acima de 0,8 para a classe 1 e entre 0,6 e 0,8 para a classe 2 (Fig. 6).

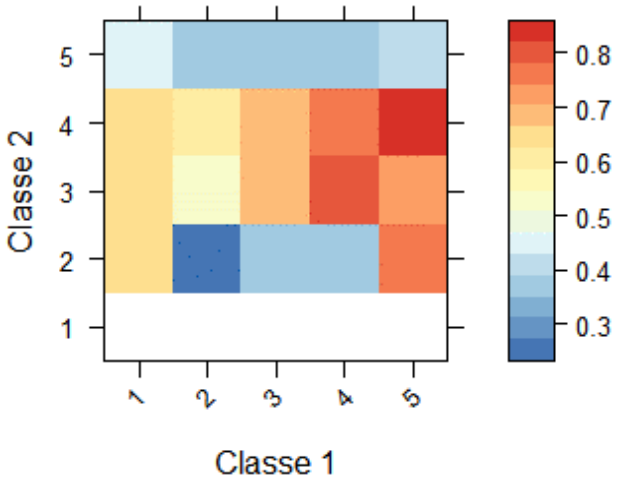

Fig. 6. Mapa de cor do Banco C (com troca de rótulo)

Novamente é mostrada a comparação entre o desempenho do método proposto e o KNN, na Tabela VII. Neste caso, onde existe erro, o método alcançou resultados bem expressivos, se saindo melhor que o classificador KNN, em todas as quantidades de observações utilizadas para a classificação.

TABELA VI. DESEMPENHO DE CLASSIFICAÇÃO NO BANCO C

	Método proposto		Seleção aleatória	
# obs	Média de desempenho	Desvio padrão	Média de desempenho	Desvio padrão
1600	**0.83**	0.00	0.81	0.00
1400	**0.83**	0.00	0.81	0.01
1200	**0.83**	0.00	0.81	0.01
1000	**0.83**	0.01	0.80	0.01
800	**0.83**	0.00	0.80	0.01
600	**0.82**	0.01	0.79	0.01
400	**0.82**	0.01	0.78	0.01
200	**0.80**	0.01	0.76	0.01
180	**0.80**	0.01	0.76	0.01
160	**0.79**	0.01	0.76	0.01
140	**0.79**	0.01	0.75	0.01
120	**0.78**	0.01	0.75	0.01
100	**0.78**	0.01	0.74	0.01
80	**0.77**	0.01	0.73	0.02
60	**0.75**	0.02	0.73	0.02
40	**0.73**	0.03	0.71	0.02
20	**0.69**	0.03	0.68	0.03
15	0.66	0.05	0.66	0.05

TABELA VII. DESEMPENHO DE CLASSIFICAÇÃO NO BANCO B COM TROCA DE RÓTULO

	Método proposto		Seleção aleatória	
# obs	Média de desempenho	Desvio padrão	Média de desempenho	Desvio padrão
600	**0.82**	0.00	0.66	0.01
400	**0.82**	0.00	0.66	0.02
200	**0.80**	0.01	0.64	0.03
180	**0.80**	0.01	0.64	0.02
160	**0.80**	0.01	0.64	0.03
140	**0.79**	0.01	0.64	0.03
120	**0.79**	0.01	0.63	0.04
100	**0.78**	0.02	0.62	0.04
80	**0.78**	0.02	0.63	0.04
60	**0.77**	0.02	0.60	0.05
40	**0.76**	0.03	0.59	0.06
20	**0.72**	0.03	0.56	0.07
15	**0.72**	0.04	0.55	0.08

IV. Conclusão e Trabalhos Futuros

Nesta pesquisa foi proposto um algoritmo de seleção de observações inspirado no classificador Bayesiano, onde as estimativas das distribuições de probabilidade foram calculadas através do uso de Kernels. Comparações realizadas com o KNN, clássico algoritmo de classificação local, foram executadas a fim de comparação do método proposto.

A seleção de observações demonstrou bons resultados ao ser utilizada com o classificador KNN. Dos 3 bancos de dados utilizados sem ruído, o método proposto alcançou uma classificação melhor em 2, se comparado ao KNN "puro". Ao se inserir a troca de rótulo, o método se saiu melhor em todos os bancos de dados.

Demonstrou-se com este artigo que através da seleção de observações é possível analisar uma área onde os dados possuem uma maior relevância para a classificação, e com isso tornar classificadores mais eficientes, além de ser um novo método de selecionar os dados mais importantes para uma determinada tarefa.

A aplicação em um problema específico de seleção de observações, e a utilização com outros classificadores além do KNN podem ser aplicações futuras para o método, para explorar todo o seu potencial.

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] D. Tuia, M. Volpi, L. Copa, M. Kanevski e J. Muñoz. "A survey of active learning algorithms for supervised remote sensing image classification" in IEEE Journal of Selected Topics in Signal Processing, vol. 5, pp. 606-617, Julho 2011.

[9] K. Wang, D. Zhang, Y. Li, R. Zhang, L. Lin. "Cost-effective active learning for deep image classification". IEEE transactions On Circuits And Systems For Video Technology, vol. 27, no. 12, Dez 2017.

[10] R. Chang, H. Hsu, S. Lin, C. Chang e J. Ho. "Query-Based Learning For Dynamic Particle Swarm Optimization". IEEE Access, vol. 5, pp. 7648 – 7658. Abr 2017.

[11] H. Khosravani, A. Ruano e P. Ferreira. "A Convex Hull-Based Data Selection Method for Data Driven Models". Applied Soft Computing. Elsevier. 2016

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

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