Análise de componentes principais e máquina de vetores de suporte com núcleo RBF aplicados a um sistema de posicionamento interior

Principal component analysis and support vector machine with RBF kernel applied to an indoor system positioning

DOI:10.34117/bjdv8n10-328

Recebimento dos originais: 26/09/2022
Aceitação para publicação: 30/10/2022

Guilherme Márcio de Melo Campos Fonte Bôa
Especialista em Aprendizado de Máquina
Instituição: Universidade do Estado de Minas Gerais
Endereço: Avenida Paraná, 3001, Jardim Belvedere I, Divinópolis - MG, CEP: 35501-170
E-mail: guimarcio93@gmail.com

Marcos Alberto Saldanha
Mestre em Engenharia Elétrica
Instituição: Universidade do Estado de Minas Gerais
Endereço: Avenida Paraná, 3001, Jardim Belvedere I, Divinópolis - MG, CEP: 35501-170
E-mail: marcos.saldanha@uemg.br

Edwaldo Soares Rodrigues
Mestre em Ciência da Computação
Instituição: Universidade do Estado de Minas Gerais
Endereço: Avenida Paraná, 3001, Jardim Belvedere I, Divinópolis - MG, CEP: 35501-170
E-mail: edwaldo.rodrigues@uemg.br

RESUMO

O Global Positioning System (GPS), que em português significa sistema de posicionamento global, tem como intuito informar a um dispositivo móvel sua localização em qualquer ponto do planeta. Apesar disso, estes sistemas possuem baixo desempenho quando os receptores estão no interior de locais fechados. O problema de localização no interior de ambientes, pode ser resolvido pelos sistemas de posicionamento interior (Indoor System Positioning - IPS). Os algoritmos de classificações de padrões podem ser implementados para localização, quando se analisa a intensidade do sinal recebido (Received Signal Strength) de Wi-Fi. Para este trabalho, foi obtido um conjunto de dados do repositório da University of California Irvine (UCI), com medidas de intensidade de sete roteadores para quatro salas diferentes. Criou-se códigos em Python 3.8 no Jupyter Notebook, para o pré-processamento dos dados, para redução de dimensionalidade por Análise de Componentes Principais (PCA), e para treino e validação cruzada (5-fold) do modelo de classificação por Máquina de Vetores de Suporte (SVM) com núcleo Função de Base Radial (RBF). A projeção sobre apenas duas componentes principais, resultou na representação de 85,75% da informação do espaço de atributos em sete dimensões. O conjunto de dados foi transformado pelo PCA, padronizado, e utilizado para a etapa de treino e validação. A acurácia do método SVM com núcleo RBF foi comparada com

outros algoritmos de aprendizado de máquina, e foi superior a todos eles, com 98,42% de acerto.

Palavras-chave: sistema de posicionamento interior, análise de componentes principais, máquina de vetores de suporte, função de base radial.

ABSTRACT
Global positioning systems (GPS), aim to inform a mobile device of its location anywhere on the planet. Despite this, these systems have low performance when the receivers are indoors. The problem of locating indoor, can be solved by an indoor positioning systems (IPS). Pattern classification algorithms can be implemented for location, when analyzing Wi-Fi Received Signal Strength (RSS). A dataset was obtained from the University of California Irvine (UCI) repository, with intensity measurements from seven routers for four different rooms. Codes were created in Python 3.8 in Jupyter Notebook, for data pre-processing, for dimensionality reduction by Principal Component Analysis (PCA), and for training and cross-validation (5-fold) of the classification model by Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel. The projection on only two principal components resulted in the representation of 85.75% of the attribute space information, that were in seven dimensions. The dataset was transformed by the PCA, standardized, and used for the training and validation stage. The accuracy of the SVM method with RBF kernel was compared with other machine learning algorithms, and was superior to all of them, with 98.42% accuracy.

Keywords: indoor positioning system, principal component analysis, support vector machines, radial basis function.

1 INTRODUÇÃO

Os sistemas de posicionamento global (GPS), têm como intuito informar a um dispositivo móvel sua localização em qualquer ponto do planeta, e isto é feito mediante a combinação de informações providas por vários satélites[9]. No entanto, estes sistemas possuem baixo desempenho quando os receptores estão no interior de casas, apartamentos ou em outros locais fechados. Construções e outras estruturas atenuam os sinais emitidos pelos satélites, o que diminui a precisão[4].

O problema de localização no interior de ambientes, o qual o GPS apresenta falhas, pode ser resolvido pelos sistemas de posicionamento interior (IPS)[1]. Existem várias abordagens que podem ser utilizadas nestes sistemas, como: sinais de rede de Wi-Fi (WPS), de Bluetooth, de radiação infravermelha e de ultrassom. A abordagem a ser escolhida, pode depender da aplicação do IPS[3].

A popularização crescente dos *smartphones* trouxe ao IPS aplicações em vários segmentos, como: acessibilidade e navegação em ambientes inteligentes (casas, edifícios, *shoppings,* aeroportos, hotéis, hospitais, museus), realidade aumentada, eventos e feiras,

prevenção de roubos (ambientes restritos), e até mesmo no gerenciamento de recursos (materiais e humanos) com objetivo na redução de custos[3,5,6,7].

O uso de IPS por Wi-Fi leva a dois grupos de métodos: os baseados em modelos e os baseados em impressão digital de rádio frequência. No primeiro, utiliza-se processamento de sinais e análise geométrica para calcular o posicionamento. Já no segundo, são obtidos indicadores de força do sinal recebido (RSS) para cada posição, e aplica-se algoritmos de classificação de padrões, a fim de obter a localização[3,8].

Neste estudo foi utilizado um conjunto de dados pré-processados, com as intensidades de sinais recebidos de sete roteadores, do repositório da *University of California Irvine* (UCI)[11]. Assim, o objetivo do trabalho consistiu em aplicar a Análise de Componentes Principais (PCA) para redução de dimensionalidade, e a Máquina de Vetores de Suporte (SVM) com núcleo de Função de Base Radial (RBF), para localização de *smartphones* de forma *offline*, dentre quatro salas diferentes. A acurácia do SVM com núcleo RBF foi comparada com alguns outros métodos de classificação.

2 REFERENCIAL TEÓRICO

2.1 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

O PCA é um procedimento de transformação linear ortogonal, o qual descorrelaciona um conjunto de dados correlacionados. Isso resulta em um novo sistema de coordenadas, com dimensão igual ou menor que o original, em que cada eixo coordenado é uma componente principal diferente[12].

Suponha um conjunto de dados na forma matricial $\boldsymbol{X}_{n\,x\,m}$, em que n representa o número de instâncias (observações), e m representa o número de atributos para cada observação. Uma matriz de transformação linear ortogonal $\boldsymbol{um}_{\,x\,k}$, em que k é o número de componentes principais, transforma $\boldsymbol{X}$ em $\boldsymbol{Y}_{n\,x\,k,}$ conforme equação (1):

$$\mathbf{Y} = \mathbf{XU} \qquad (1)$$

A matriz de transformação $\boldsymbol{U}$ pode ser calculada pela decomposição espectral da matriz de covariância $\boldsymbol{C_X}$, do conjunto de dados, o que é mostrado na equação (2):

$$\mathbf{C_X} = \mathbf{S\Lambda S^{-1}} \qquad (2)$$

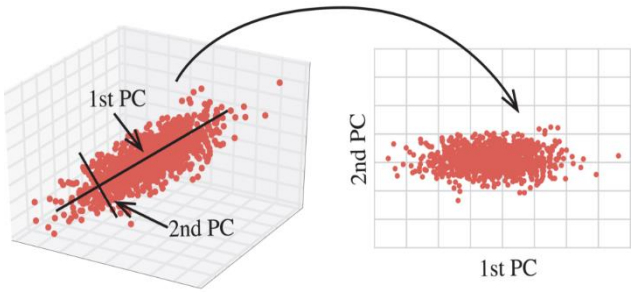
em que **S** é a matriz de autovetores, e **Λ** é a matriz diagonal de autovalores.

Os autovalores de $\boldsymbol{C_X}$ representam as variâncias nas direções dos autovetores (componentes principais). Projetar os dados sobre autovetores de pequenos autovalores, representa perder grande parte da variância dos dados. Assim, neste processo de compressão de dados, é importante projetar sobre os autovetores dos maiores autovalores, de forma que haja o mínimo de perda de informação possível. Portanto, suponha os autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$, e que se deseja projetar os dados originais sobre k componentes principais, de forma que $k \leq m$. Assim, a matriz de transformação $\boldsymbol{U}$ pode ser escrita em termos dos seguintes autovetores, como mostrado em (3):

$$\mathbf{U} = \begin{bmatrix} \mathbf{s_1} & \mathbf{s_2} & \ldots & \mathbf{s_k} \end{bmatrix} \qquad (3)$$

A Figura 1 apresenta de forma intuitiva e gráfica o processo de redução de dimensionalidade pelo PCA. Neste caso, o conjunto de dados original em três dimensões, foi transformado em um novo conjunto de dados descorrelacionados, com duas dimensões, com eixos representados pelo primeiro e segundo componente principal.

Figura 1 - Redução de dimensão e descorrelação de um conjunto de dados pelo PCA.



Fonte: Medium, 2022[13].

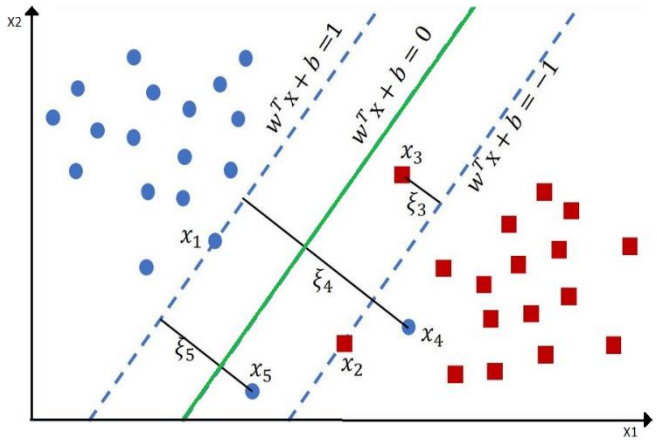
2.2 MÁQUINA DE VETORES DE SUPORTE (SVM)

O SVM é uma técnica de aprendizado supervisionado, em que se busca encontrar um hiperplano com a margem maximizada, que será o limite de decisão e será aplicado na tarefa de classificação. A margem é definida pelos vetores de suporte, que passam por pontos de extremos para uma dada classe. Quando o conjunto de dados não for linearmente separável, admite-se uma variável de folga, a qual permite violar algumas restrições. Esta variável de folga ξ, é a distância entre o ponto que está fora da região delimitada pelos vetores de suporte e o vetor de suporte de sua classe, como mostra a

Figura 2. Portanto, o problema primal de otimização para um SVM com duas classes é mostrado em (4)[15,16,17]:

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ s.a: \quad & y_i(w^T x_i + b) - 1 + \xi_i \geq 0 \quad \forall \quad i \in \{1, ..., n\} \end{aligned} \tag{4}$$

Figura 2 – Classificação binária de padrões pelo SVM. Fronteira de decisão de verde e as retas de suporte (tracejadas em azul). A distância entre as retas de suporte equivale a $2\|\mathbf{w}\|^{-1}$.



Fonte: Modificado de ResearchGate, 2022[16].

O hiperparâmetro **C** é o termo regularizador, que pondera entre margem e folga. Quando C □ ∞, a restrição da folga será mais forte e a margem será menor. Se o valor de C □ 0, permitirá uma margem maior, porém com mais pontos classificados de forma errada. Este termo pode ser escolhido por *grid search*, o qual resulta em um modelo para cada valor do hiperparâmetro.

O problema de otimização dual de (1) é dado por (5)[17]:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_i \alpha_i \\ s.a: \quad & \sum_i \alpha_i y_i = 0, \quad \alpha \in [0, C] \end{aligned} \tag{5}$$

Ao resolver o problema dual, pode-se encontrar a expressão de classificação de uma nova observação *x*, dada por (6)[17]:

$$f(x) = \sum_i \alpha_i y_i K(x_i, x) \quad (6)$$

A função $K(x_i,x_j)$, é chamada de núcleo ou *kernel*. Esta função pode ser linear, polinomial, sigmoide ou de base radial. Ela é responsável por mapear a observação para um espaço de dimensão superior, de forma que os dados que não são linearmente separáveis no espaço dos atributos possam ser separados de forma linear, com menor erro, neste espaço de alta dimensionalidade. Apesar da fronteira de decisão no espaço mapeado ser linear, no espaço dos atributos os limites de decisão serão não lineares (se a função também for)[17]. Dentre as funções mencionadas, a função de base radial é amplamente aplicada e leva a resultados satisfatórios[2]. O núcleo RBF é definido por (7):

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2} \quad (7)$$

O termo γ é um hiperparâmetro que deve ser escolhido via *grid search,* durante a validação do modelo.

Para o caso de múltiplas classes, uma abordagem é utilizar o método um-contra-um. Supondo *k* classes, encontra-se um classificador que separe duas classes de cada vez. Portanto o número *J* de classificadores é dado por (8)[15]:

$$J = \frac{k(k-1)}{2} \quad (8)$$

A combinação de todos os J classificadores, resulta em um classificador de múltiplas classes.

3 METODOLOGIA

A realização do presente estudo se deu pelas etapas descritas detalhadamente nas subseções sequentes. Na Subseção 3.1, é apresentada a etapa de pré-processamento. A Subseção 3.2 detalha como o PCA é aplicado no conjunto de dados para reduzir a dimensionalidade. Por fim, na Subseção 3.3, é explicitado como os modelos são treinados

e validados. Todas as etapas foram realizadas no Jupyter Notebook, com os códigos escritos em Python 3.8.

3.1 PRÉ-PROCESSAMENTO

A primeira etapa foi a de pré processar os dados. O conjunto de dados consiste de uma matriz de 2000 linhas e 8 colunas. Cada linha representa uma medição da força do sinal recebido de cada um dos sete roteadores, em uma determinada sala. Ao todo, foram realizadas 500 medições para cada sala. Já as colunas, são os valores de intensidade em decibel mW (dBmW) de cada um dos roteadores. A última coluna informa qual é o rótulo de classe, neste caso, inteiros de 1 a 4, os quais representam sala 1, sala 2, e assim sucessivamente.

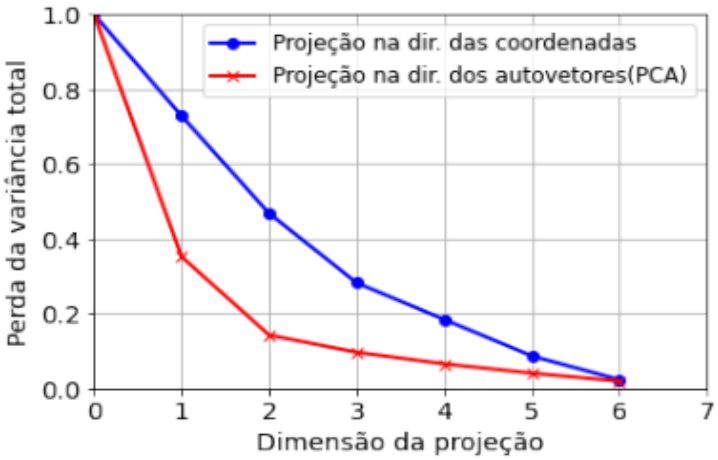
O pré-processamento consistiu em alterar o intervalo dos rótulos de classe de 0 a 3 para facilitar a escrita do código, analisar valores ausentes e instâncias duplicadas, além de filtrar as anomalias pelo método *Isolation Forest,* o qual possui ótimos resultados em conjuntos de alta dimensionalidade[18]. Este método já está embutido no pacote *SciKitLearn,* e foi aplicado com os seguintes parâmetros da função: *max_samples=500, random_state=1, contamination='auto', max_features=7.* O primeiro e último parâmetro, está relacionado com a quantidade de amostras e atributos totais, respectivamente, que serão utilizados no treinamento da floresta. O segundo, diz respeito à pseudo aleatoriedade durante a criação das árvores e foi escolhido o valor "1" para a semente aleatória. Já o terceiro parâmetro, representa a fração de contaminação dos dados por anomalias. Como cada classe poderia ter quantidades diferentes de *outliers*, o valor escolhido foi 'auto'.

Após todo o tratamento de dados descritos anteriormente, restaram 1838 linhas na matriz de dados.

3.2 REDUÇÃO DE DIMENSIONALIDADE

A redução de dimensionalidade pelo PCA, reduz o tempo computacional de algoritmos de aprendizado de máquina[10]. Todo o processo descrito na subseção 2.1 foi aplicado neste conjunto de dados. A Figura 3 apresenta a perda de informação de 0 a 1 (0% a 100%) pelo número de dimensões. Além disso, projeções feitas nas coordenadas do espaço dos atributos também foi realizada, para comparações.

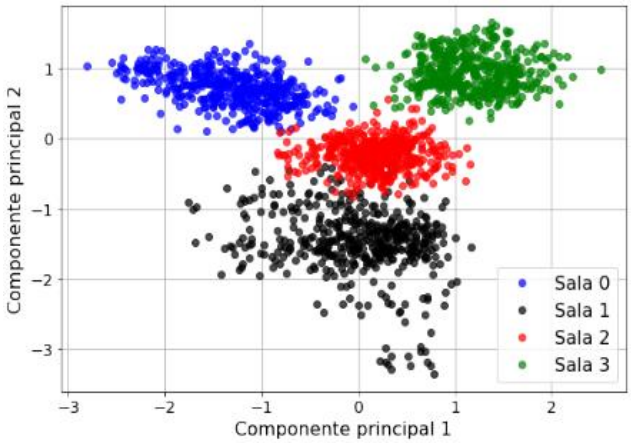
Figura 3: Perda de informação de acordo com a dimensionalidade do conjunto de dados. De azul, a projeção sobre as coordenadas do espaço de atributos. De vermelho, projeção sobre os autovetores da matriz de covariância dos dados.



Observa-se que a projeção sobre as coordenadas do espaço dos atributos (azul), proporciona menor ganho de informação com o acréscimo de dimensões, em relação ao método PCA (vermelho). Neste estudo, utilizou-se apenas duas componentes principais, já que elas conseguiram representar 85,75% de toda a informação, de um espaço de 7 dimensões.

Por ser um conjunto de dados pequeno, com treinamento e validação *offline*, a redução foi realizada antes do treino, e todo o conjunto de dados reduzido foi padronizado e armazenado. O diagrama de dispersão dos dados transformados pelo PCA e padronizados, é apresentado na Figura 4.

Figura 4: Diagrama de dispersão de todo o conjunto de dados transformado pelo PCA e padronizado.



3.3 TREINO E VALIDAÇÃO

As etapas de treino e validação foram realizadas simultaneamente, já que o conjunto de dados é único e pequeno, não havendo distinção de dados para treino e para teste. Nesta etapa, o conjunto de dados utilizado foi o reduzido pelo PCA e padronizado.

A biblioteca *SciKitLearn* possui o algoritmo do SVM com núcleo RBF já incluso, e ele foi utilizado neste trabalho. O método de treino e validação do modelo foi por meio da validação cruzada por 5-*fold*, com 80% dos dados destinados ao treino, e o restante para a validação. O *grid search* auxiliou na escolha dos valores dos hiperparâmetros, e permitiu obter um modelo com ótima generalização, que não estivesse superajustado aos dados de treino. Por fim, aplicou-se estas mesmas etapas a outros quatro métodos de aprendizado de máquina, para comparação da acurácia neste conjunto de dados, os quais foram: k Vizinhos Próximos (kNN), *Naive* Bayes, Árvores de Decisão e SVM com núcleo linear.

4 RESULTADOS E DISCUSSÕES

Os resultados de acurácia dos cinco modelos obtidos, podem ser resumidos pela Tabela 1.

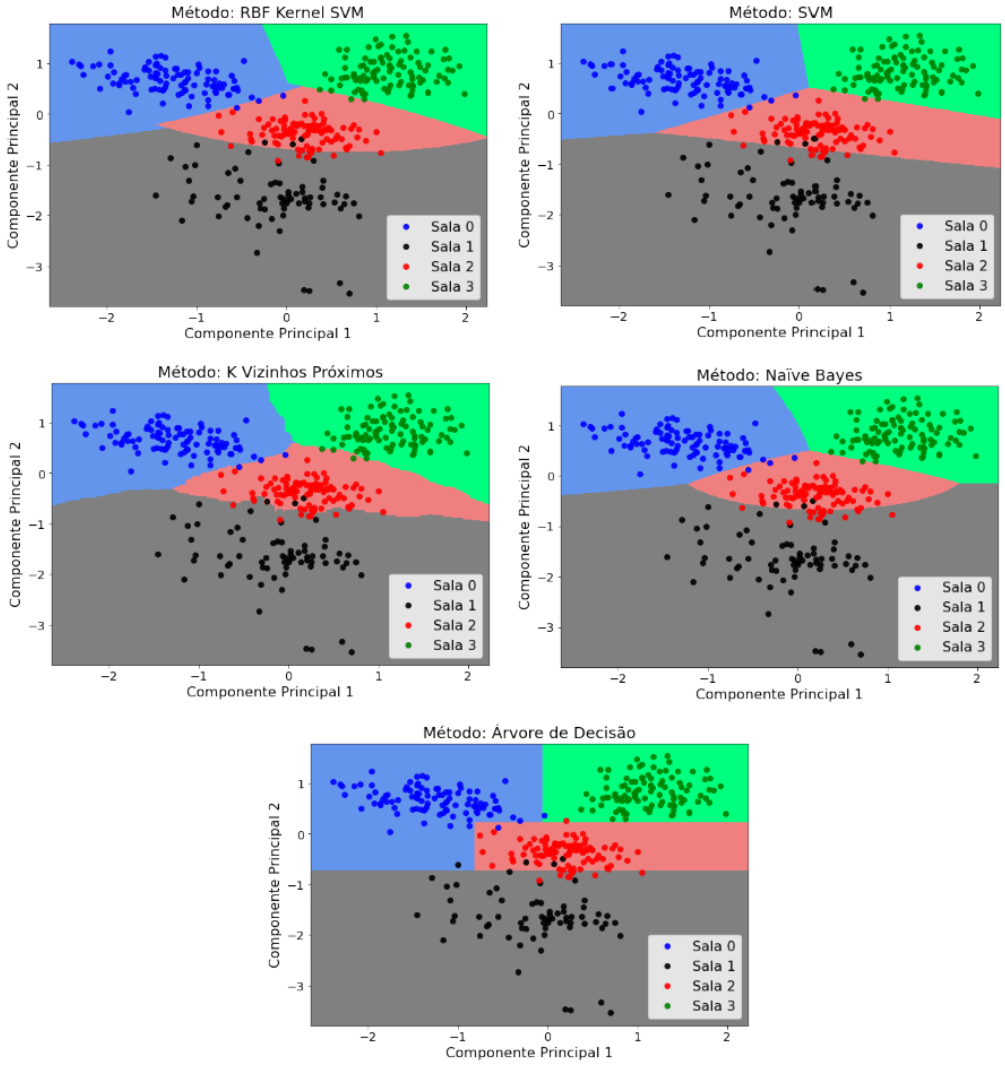
Tabela 1: Resultados comparativos das acurácias do SVM com núcleo RBF, com a acurácia de outros métodos, com seus respectivos hiperparâmetros.

Método	Acurácia: Treino	Acurácia: Validação	Hiperparâmetros
SVM núcleo RBF	98,53%	98,42%	C=10 γ=0.1
SVM núcleo Linear	98,52%	98,26%	C=2.5
kNN	98,60%	98,09%	k = 9
***Naive* Bayes**	98,55%	97,87%	-
Árvores de Decisão	98,26%	97,27%	'*criterion*' = gini '*max_depth*' = 3 '*max_features*' = 2 '*min_samples_leaf*' =10

Como pode ser observado pela tabela, todos os modelos possuem acurácia superior a 97,2% durante a validação do conjunto de dados reduzido em dimensão pelo PCA, mesmo com a perda de 14,25% da informação total. Outro fator importante a ser analisado é a capacidade de generalização dos modelos. Os valores de acurácia durante o treino e durante a validação são próximos, portanto, os modelos não sofrem de

superajustamento, possuindo assim ótima capacidade de generalização. Por fim, observa-se que mesmo com fronteiras de decisão diferentes como mostra na Figura 5, todos os métodos obtiveram acurácias de validação próximas. Dentre os métodos, a melhor acurácia obtida foi do método SVM com núcleo RBF. Isso mostra que fronteiras de decisão não lineares para o SVM, podem resultar em melhores resultados para dados linearmente não separáveis.

Figura 5 – Fronteiras de decisão de cada método, formadas pelo conjunto de treino, durante uma das iterações da validação cruzada 5-*fold*. Os pontos representam os dados de teste.



5 CONCLUSÃO

Neste trabalho, apresentou-se os sistemas de posicionamento interior, e foi aplicado a um conjunto de dados de intensidades de Wi-Fi, o método PCA para redução de dimensionalidade com menor perda de informação, e o método SVM com núcleo RBF para localização.

No pré-processamento, a filtragem de anomalias dos dados foi efetuada pelo método *Isolation Forest*. Depois, a redução de dimensionalidade do conjunto de dados foi realizada pelo PCA. Finalmente, o conjunto de dados resultante foi utilizado para treino e validação do classificador.

Concluiu-se que pelo PCA, a perda de informação é menor quando os dados são projetados sobre os autovetores da matriz de covariância, ao contrário da projeção sobre as coordenadas do espaço dos atributos. Com apenas dois componentes principais, foi possível obter 85,75% de toda informação retida no espaço dos atributos com sete dimensões.

O método SVM com núcleo RBF permitiu formar fronteiras de decisão não lineares, e isso impactou na acurácia do modelo, já que este método obteve o maior número de acertos na etapa de validação (98,42%), comparado a outros métodos de aprendizado de máquina.

Portanto, conclui-se que pelo elevado valor de acurácia, o SVM com núcleo RBF e o PCA podem ser aplicados em sistema de posicionamento interior, baseados em impressão digital de rádio frequência.

REFERÊNCIAS

KAEMARUNGSI, Kamol; KRISHNAMURTHY, Prashant. Modeling of Indoor Positioning System Based on Location Fingerprinting. Pittsburgh: IEEE, 2004.

PRAJAPATI, Gend L.; PATLE, Arti. On Performing Classification Using SVM with Radial Basis and Polynomial Kernel Functions. Goa: IEEE, 2010.

ROHRA, Jayant G. *et al.* User Location in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithm with Neural Networks. Patiala: Springer, 2017.

CHANG, T. Y.; CHIEN, Y. R. Indoor Positioning Method for Smart Mobile Device Based on Fuzzy Wi-Fi Fingerprint. *IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 2020.

HERRERA, J. C. A. *et al.* Pedestrian indoor positioning using smartphone multi-sensing, radio beacons, user positions probability map and IndoorOSM floor plan representation. *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2014.

JOANNE, W. *et al*. Indoor navigation and localisation application system. *3rd International Conference on Electronic Design (ICED)*, 2016.

YAO, W.; MA, L. Research and Application of Indoor Positioning Method Based on Fixed Infrared Beacon. *37th Chinese Control Conference (CCC)*, 2018.

YAZICI, A.; KESER, S. B.; GUNAL, S. Integration of classification algorithms for indoor positioning system. *International Conference on Computer Science and Engineering (UBMK)*, 2017.

VAN DIGGELEN, F. *A-GPS: Assisted GPS, GNSS, and SBAS*, Artech, 2009.

SALAMAH, A. H. *et al.* An enhanced WiFi indoor localization system based on machine learning. *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2016.

UCI, Disponível em:https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization, 2017. Acessado dia 03/08/2022.

JOLLIFFE, I. T. Principal Component Analysis, 2ª ed. Springer, 2002.

VADALI, S. G. DAY 10: Dimensionality Reduction with PCA and t-SNE in R. Medium. Disponível em https://medium.com/@TheDataGyan/dimensionality-reduction-with-pca-and-t-sne-in-r-2715683819. Acessado dia 04/08/2022.

KOMURA, D. *et al.* Multidimensional support vector machines for visualization of gene expression data. *Bioinformatics*, 2005.

WESTON, J.; WATKINS, C. Multi-class Support Vector Machines. Technical Report. University of London, 1998.

HUNG, L.; TRAN, T.; LANG, T. Automatic Heart Disease Prediction Using Feature Selection and Data Mining Technique. Journal of Computer Science and Cybernetics. 2018.

BISHOP, C. M. Pattern recognition and machine learning. Springer, 2006.

LIU, F. T.; TING, K. M.; ZHOU, Z. Isolation Forest. *Eighth IEEE International Conference on Data Mining, 2008.*