INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO SUL DE MINAS GERAIS
CÂMPUS MUZAMBINHO
Curso de Ciência da Computação

Rafael Vicente da Silva

Avaliação da Classificação de Imagens Hiperespectrais Após a Redução da Dimensionalidade Usando Algoritmos de *Manifold Learning*

Muzambinho

2021

Rafael Vicente da Silva

Avaliação da Classificação de Imagens Hiperespectrais Após a Redução da Dimensionalidade Usando Algoritmos de *Manifold Learning*

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação, do Instituto Federal de Educação Ciência e Tecnologia do Sul de Minas Gerais - Câmpus Muzambinho, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientador: Diego Saqui

Muzambinho

2021

COMISSÃO EXAMINADORA

Prof. Dr. Diego Saqui (Orientador)

Prof 1.: XXXXXXX

Prof 1.: XXXXXXX

Prof 1.: XXXXXXX

Muzambinho, ____ de _____ de 2021

SILVA, Rafael Vicente da. <Avaliação da Classificação de Imagens Hiperespectrais Após a Redução da Dimensionalidade Usando Algoritmos de *Manifold Learning*>. 2021. <xx. E: 65f>. Trabalho de Conclusão de Curso (Curso Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Muzambinho, Muzambinho, 2021.

RESUMO

Imagens hiperespectrais (IHs) possuem centenas de bandas e maior capacidade de discriminação de diferenças sutis em comparação a imagens multiespectrais, oferecendo informações quimicas e fisiológicas do local, o que beneficia aplicações de precisão. Entretanto, a alta resolução espectral e alta correlação de bandas inerentes dessas imagens, sugerem a possibilidade de ocorrência da maldição da dimensionalidade (HUGHES, 1968) em processos de reconhecimento de padrões. Dessa forma, o estudo dos efeitos de métodos de redução de dimensionalidade (RD) é relevante para esse tipo de imagem. Adicionalmente, é fundamental a comparação de comportamento de métodos tradicionais, como o PCA (Análise de Componentes Principais), com métodos de manifold learning, como o t-SNE (t-distributed Stochastic Neighbor Embedding). Nesse cenário, o objetivo do presente trabalho é analisar a classificação final da imagem hiperespectral, Indian Pines, obtida pelo sensor da NASA, AVIRIS (Airborne Visible / Infrared Imaging Spectrometer), fazendo uso dos classificadores KNN (*K - Nearest Neighbors*) e SVM (*Support Vector Machine*) posteriormente ao processo de RD a partir de algoritmos da classe de Manifold Learning.

Palavras-chave: extração de característica, manifold learning, redução de dimensionalidade, imagens hiperespectrais.

ABSTRACT

Hyperspectral images (HIs) have hundreds of bands and greater ability to discriminate subtle differences compared to multispectral images, offering chemical and physiological information about the site, which benefits precision applications. However, the high spectral resolution and high correlation of bands inherent in these images, suggest the possibility of the occurrence of the curse of dimensionality (HUGHES, 1968) in pattern recognition processes. Thus, the study of the effects of dimensionality reduction (RD) methods is relevant for this type of image. Additionally, it is relevant to compare the behavior of traditional methods, such as PCA (Principal Component Analysis), with manifold learning methods, such as TSNE (t-distributed Stochastic Neighbor Embedding). In this scenario, the objective of the present work is to analyze the final classification of hyperspectral images obtained by the NASA sensor, AVIRIS (Airborne Visible / Infrared Imaging Spectrometer), using the KNN (K - Nearest Neighbors) and SVM (Support Vector Machine) classifiers. after the RD process using algorithms of the Manifold Learning class.

Keywords: feature extraction, manifold learning, dimensionality reduction, hyperspectral images.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. Introdução

O conceito de *Big Data* está relacionado à capacidade de processar e analisar grandes volumes de informação que permitam a extração de conhecimentos úteis para melhorar o processo de tomada de decisão (MARQUESONE, 2016). Com isso, os grandes volumes de dados podem ser caracterizados em relação a sua quantidade e também à quantidade de seus atributos (dimensões) (JACOBS, 2009; LAZER et al., 2014). Dessa forma, surgem técnicas envolvendo o desenvolvimento e aplicação de métodos de reconhecimento de padrões e mineração de dados (DISNER, 2015). Exemplos de sistemas que operam com dados de alta dimensão e/ou *Big Data* incluem áreas como análise de dados geoespaciais, bioinformática, organização e recuperação de imagens baseadas em conteúdo, bases de dados distribuídas na internet, redes de sensores e imagens hiperespectrais (IHs).

Em especial, as IHs ilustram a composição química por meio de imagens feitas a partir de informações espectrais coletadas por um espectrômetro (Figura 1), cujo sensor hiperespectral obtém milhares ou centenas de milhares de espectros, ao invés de um único espectro (GHAMISI et al., 2017). Pode-se mostrar que classes espectralmente muito semelhantes, isto é, classes que compartilham de vetores de médias, muito próximos entre si, podem ser separadas com um grau de acurácia em espaços de dimensão suficientemente alta (FUKUNAGA, 1990).

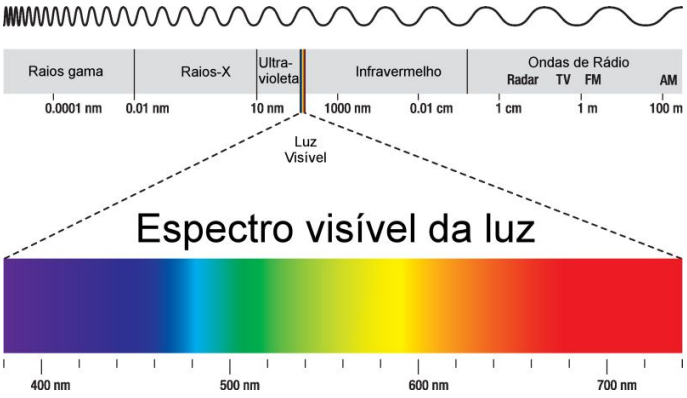

Figura 1: Comportamento do comprimento de onda de espectro.

Fonte: Info Escola

Em razão do avanço da computação e dos sistemas sensores, surgiram novas possibilidades de manipulação no domínio espectral por meio do Sensoriamento Remoto (SR). Segundo Florenzano (2007), SR é definido como a tecnologia que possibilita obter imagens - e outros tipos de dados - da superfície

terrestre, por meio da captação e do registro da energia refletida ou emitida pela superfície. O termo sensoriamento está relacionado à obtenção dos dados por meio de sensores situados em plataformas terrestres, aéreas e orbitais, e o termo remoto, que significa distante, é utilizado pois a obtenção é feita sem o contato físico do sensor ao objeto de estudo, como ilustrado na Figura 2.

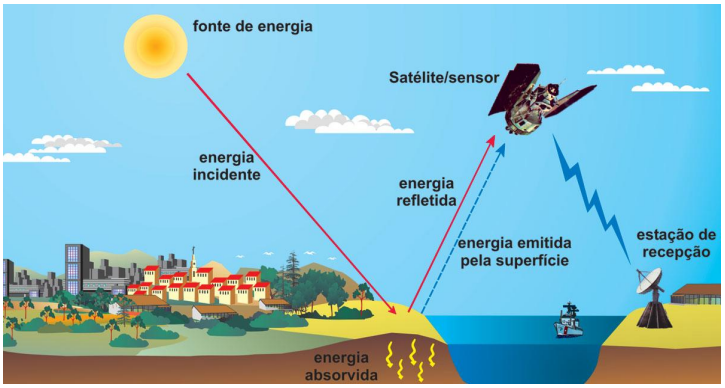

Figura 2: Obtenção de imagens por SR.

Fonte: FLORENZANO, 2007.

Utilizando sensores de alta resolução espectral, que proporcionam para cada pixel (elemento de resolução espacial), medidas radiométricas em bandas estreitas e contínuas, pode-se obter uma grande quantidade de informações espectrais em seu domínio. Essas informações têm um nível de resolução mais próximo daquele verificado em espectrorradiômetros de campo ou de laboratório, facilitando o uso de abordagens mais específicas, que permitam quantificar alvos com maior nível de detalhamento espectral, compondo assim as IHs (CLARK, 1999).

IHs geralmente são vistas como cubos hiperespectrais, onde as imagens de banda simples estão empilhadas de modo que a terceira dimensão do cubo é incrementada pelos comprimentos de onda amostrados, como demonstra a Figura 3.

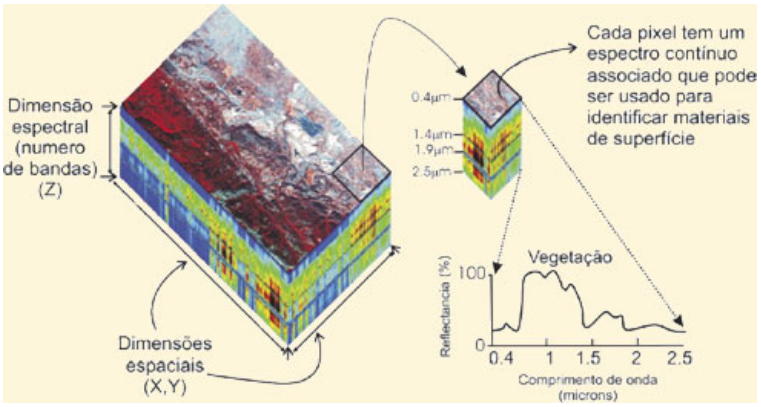

Figura 3: Ilustração do comportamento de IHs.

Fonte: MUNDOGEO, 2021.

Uma IH que é coletada em d bandas espectrais pode ser pensada como uma nuvem de pontos (dados pelos pixels na imagem) em um espaço d-dimensional. As d bandas espectrais na imagem formam d eixos de coordenadas no hiperespaço. Cada pixel é representado como um vetor d-dimensional tal que o valor da i-*ésima* coordenada é o valor estimado da reflectância terrestre (ou da radiância do alcance do sensor) medida no i-*ésimo* comprimento de onda. Esse vetor é a representação geométrica do pixel no espaço espectral.

A complexidade e o grande volume de dados inerentes, exigem software adequado para a sua análise (ZHANG et al., 2000) e a utilização de algoritmos de classificação apropriados. Entre eles pode-se citar os sensores *Airbone Visible Infrared Imager Spectroradiometer* ("*AVIRIS*"), com 224 bandas espectrais e o *HYPERION* a borda do satélite *Hyperspectral Imager (Earth Observing 1)*, com 220 bandas espectrais. Uma dessas características descreve a necessidade de uma elevada quantidade de pixels rotulados para amostras de treinamento, que normalmente são apresentados por meio de mapas de Ground Truths (GTs).

Jimenez et al. (1998) apontam que o impacto do problema da complexidade da dimensionalidade dos dados varia de um campo para outro. Para a otimização combinatória de muitas dimensões é visto como um crescimento exponencial do esforço computacional, já no campo estatístico se manifesta como um problema de parâmetro ou densidade estimativa, devido à escassez de dados.

Os dados/registros computacionais que possuem uma quantidade elevada de atributos (dimensões), dizemos que são dados superdimensionados. Dados superdimensionados oferecem um poder discriminante mais elevado do que dados com baixa dimensionalidade, contudo a análise destes pode evidenciar um problema chamado de maldição da dimensionalidade ou fenômeno de Hughes (HUGHES, 1968).

A maldição da dimensionalidade é um fator desafiador na modelagem matemática, visto que para um hiperplano cartesiano com d dimensões de entrada onde cada dimensão de entrada é particionada em s^d células, o número total de células seria s^d (BELLMAN, 1961). Como consequência disso, a criação de modelos destes dados necessita considerar espaços de busca inerentemente esparsos (LAROSE, 2006). Dessa forma, os cientistas têm-se deparado com a necessidade de encontrar estruturas significativas ocultas de baixa dimensão, dentro de dados de

alta dimensão, sendo tal técnica denominada de redução de dimensionalidade dos dados (RDD), (PEARSON, 1901; HOTELLING, 1933; JOLLIFE, 2003; ROWEIS et al., 2003; DONOHO et al., 2003).

O efeito negativo dessa escassez resulta de alguma geometria, estatística e propriedades assintóticas do espaço de recursos de alta dimensão, e essas características exibem um comportamento surpreendente para dados em dimensões superiores. Os autores Kendall (1961) e Scott (1992) descrevem que a concentração de dados no hiperespaço à medida que a dimensionalidade aumenta, tendem a ficarem isolados em determinados pontos do hiperespaço.

A partir disso, devido ao grande volume de dados (bandas espectrais de IHs), a extração dessas informações não é uma tarefa trivial, onde são necessários o uso de teorias e ferramentas para o auxílio na extração e análise de informações úteis, facilitando assim a classificação (BORGES et al., 2006). Porém, para tratar de dados com alta dimensionalidade, podemos utilizar estratégias baseadas em *Manifold Learning* (ML). ML partem essencialmente da distância (correlação) que os dados estão dispostos entre si no espaço, o grau de afinidade nos dados, e demais métricas irão reduzir a dimensionalidade da base de dados e simultaneamente preservar a relação dos mesmos (CAYTON, 2005).

Considerando o contexto anterior, este trabalho pretendeu fazer o uso de algoritmo de ML, para a redução da dimensionalidade dos dados (bandas espectrais de uma IHs) de modo a obter uma base de dados simplificada, facilitando dessa maneira a extração de informações a respeito da mesma, possibilitando que os classificadores obtenham resultados mais assertivos de acurácia, ao qual o propósito principal foi utilizar ML, para posteriomente otimizar a acurácia de classificação de uma IHs, demonstrando assim qual algoritmo terá melhor uso para RD aos que necessitarem da mesma, como pesquisadores de Estatísticas, *Machine Learning,* Ciência da Computação e Desenvolvedores destes campos de estudo.

2. Objetivos

2.1. Objetivo Geral

Avaliar a redução da dimensionalidade dos dados em imagens hiperespectrais (IHs) a partir do uso de algoritmos baseados em *Manifold Learning* para o propósito de classificação de IHs. Com isso explorar o problema da maldição da dimensionalidade, verificando o comportamento dos algoritmos empregados,

investigando o que melhor aplicou-se obtendo mais precisão na classificação dos pixels da base de dados específica de estudo, Indian Pines, analisando a precisão do resultado da classificação, de modo a investigar a eficiência.

2.2. Objetivos Específicos

- Aplicar os redutores Análise de Componentes Principais (PCA) e t-distributed Stochastic Neighbor Embedding (t-SNE), a fim de identificar o qual melhor aplica-se na base de Indian Pines.
- Realizar comparações entre os algoritmos redutores utilizados, a fim de identificar qual melhor aplicou-se na base de Indian Pines.
- Aplicar os classificadores K-Nearest Neighbors (KNN) e Support Vector Machines (SVM), a fim de identificar qual melhor aplica-se à base de Indian Pines .
- Realizar comparações entre os classificadores, de modo a apurar qual obterá melhor resultado classificando a IHs de Indian Pines.
- Gerar uma arquitetura que representou o procedimento completo utilizado neste estudo.

3. Revisão de literatura

Este capítulo apresenta os principais conceitos que foram utilizados neste projeto e trabalhos relacionados. Inicialmente, é destacado o comportamento dos dados no hiperespaço. Na sequência têm-se os algoritmos a serem empregados, sendo dois redutores, análise de componentes principais, do inglês *Principal Components Analytics* (PCA), é uma técnica multivariada de modelagem da estrutura de covariância e Incorporação de vizinhos estocásticos com distribuição t, do inglês *t-Distributed Stochastic Neighbor Embedding* (t-SNE), na qual converte as distâncias entre os pontos no espaço multidimensional em probabilidades que representam as similaridades, e dois classificadores, K*-ésimo* vizinho mais próximo, do inglês *K-Nearest Neighbors* (KNN), cuja a classificação é feita por meio da busca dos k vizinhos, utilizando uma medida de distância nesta procura e suporte de máquina vetorial, do inglês *Support Vector Machine* (SVM), têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes.

3.1 Demonstração matemática sobre a dimensionalidade dos dados

Nesta seção, ilustramos alguns fatos incomuns ou inesperadas características do hiperespaço (espaços para mais de três dimensões), incluindo uma prova e discussão. Tais ilustrações pretendem mostrar que o espaço dimensional superior é bastante diferente do espaço 3-D com o qual nós somos familiares.

Conforme a dimensionalidade aumenta:

A. O volume de um hipercubo concentrado nos cantos, conforme ilustra na Figura 4 (SCOTT, 1992)

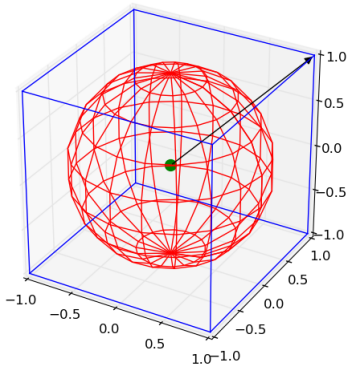

Figura 4: Comportamento dos dados no hiperespaço.

Fonte: Do autor

Foi demonstrado (KENDALL, 1961) que o volume da hiperesfera de raio r e dimensão d é dado por:

$$V_s(r) = volume\ da\ hiperesfera = \frac{2r^d}{d}\frac{\pi^{d/2}}{\Gamma(d/2)} \qquad (1)$$

e que o volume de um hipercubo em $[-r, r]^d$ é dado por:

$$V_c(r) = volume\ do\ hipercubo = (2r)^d \qquad (2)$$

A fração do volume de uma hiperesfera inscrita em um hipercubo é:

$$f_{d1} = \frac{V_s(r)}{V_c(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \qquad (3)$$

onde d é o número de dimensões. Vemos na Fig. 5 como f_{d1} diminui à medida que a dimensionalidade aumenta.

Observe que $lim_{d\to\infty} f_{d1} = 0$, o que implica que o volume do hipercubo está cada vez mais concentrado nos cantos conforme d aumenta.

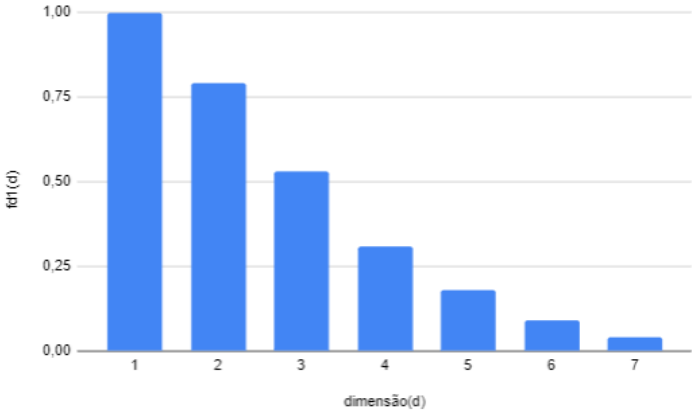

Figura 5. Volume fracionário de uma hiperesfera inscrita em um hipercubo como uma função da dimensionalidade.

Fonte: Jimenez et al. (1998)

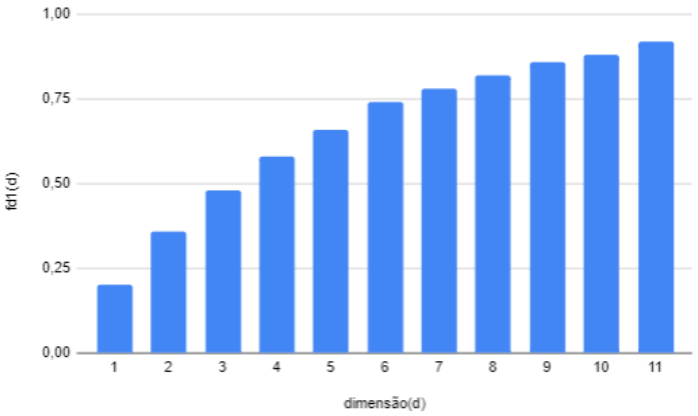

Figura 6. Volume de uma hiperesfera contida na casca externa como uma função de dimensionalidade para $\varepsilon = r/5$

Fonte: Jimenez et al. (1998)

As características mencionadas anteriormente têm duas consequências importantes para dados de alta dimensão que aparecem imediatamente. O primeiro é que o espaço de alta dimensão é quase vazio, o que implica que os dados multivariados geralmente possuem uma estrutura dimensional inferior. Como consequência, dados de alta dimensão podem ser projetados para uma dimensão inferior ao subespaço sem perder informações significativas, em termos de separabilidade entre as diferentes classes estatísticas. A segunda consequência do exposto é que os dados normalmente distribuídos terão tendência a se concentrar nas caudas; similarmente, dados uniformemente distribuídos terão maior probabilidade de serem coletados nos cantos, tornando a estimativa da densidade mais difícil. As vizinhanças locais estão quase certamente vazias, exigindo que a largura de banda de estimativa seja grande e produzindo o efeito de perder estimativa de densidade detalhada.

Suporte para esta tendência pode ser encontrado nas estatísticas, o comportamento de multivariadas normalmente são uniformemente distribuídas para dados em alta dimensionalidade. Espera-se que, à medida que a dimensionalidade aumente, os dados se concentrem em uma casca externa.

Dessa maneira, conclui-se que o hiperespaço de certa forma acaba sendo um “grande vazio” e os dados concentrando-se nas bordas, evidencia-se o fato da simplificação do problema usando da redução de dimensionalidade e ainda a

correlação dos dados de alta dimensão aproxima-se muito para dados de baixa dimensão.

3.2. Algoritmos empregados

3.2.1. Redutor: Análise de Componentes Principais

A análise de componentes principais, do inglês *Principal Components Analytics* (PCA), é uma técnica multivariada de modelagem da estrutura de covariância. A técnica foi inicialmente descrita por Pearson (1901) e uma descrição de métodos computacionais práticos veio muito mais tarde com Hotelling (1933, 1936) que usou com o propósito determinado de analisar as estruturas de correlação. A PCA é uma técnica estatística de análise multivariada que transforma linearmente um conjunto original de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original.

PCA é a técnica mais conhecida e está associada à ideia de redução de massa de dados, com menor perda possível da informação, contudo é importante ter uma visão conjunta de todas ou quase todas as técnicas da estatística multivariada para resolver a maioria dos problemas práticos, também é associada à ideia de redução de massa de dados, com menor perda possível da informação. Procura-se redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados (MANLY, 1986; HONGYU, 2015).

O objetivo principal da análise de componentes principais é o de explicar a estrutura da variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais, que são as dimensões, e são não correlacionadas entre si (SANDANIELO, 2008).

3.2.2. Redutor: Incorporação de vizinhos estocásticos com distribuição t

O algoritmo t-SNE (*t-Distributed Stochastic Neighbor Embedding*), da classe de *Manifold Learning* fornece um método eficaz para visualizar um conjunto complexo de dados. Ele descobre com sucesso estruturas ocultas nos dados, expondo clusters naturais e suavizando variações não-lineares ao longo das dimensões, reduzindo para duas ou três dimensões (VAN DER MAATEN, 2008).

Muitos conjuntos de dados do mundo real têm uma baixa dimensionalidade intrínseca, apesar de estarem inseridos em um espaço de alta dimensão. Esse espaço de baixa dimensão está embutido no espaço de alta dimensão de uma maneira complexa e não linear. Escondido nos dados, esta estrutura só pode ser recuperada através de métodos matemáticos específicos (ROSSANT, 2015).

O método, em síntese, converte as distâncias entre os pontos no espaço multidimensional em probabilidades que representam as similaridades. No espaço dimensional reduzido, as distâncias são também calculadas, sendo posteriormente ajustadas conforme o cálculo do gradiente, que representa a similaridade posicional dos pontos em relação a ambos os espaços dimensionais.

3.2.3. Classificador: *K - Nearest Neighbors*

O KNN (*K-Nearest Neighbors*) é um dos algoritmos de classificação mais utilizados na área de aprendizagem de máquina (DINIZ et., 2013). É baseado na procura dos k vizinhos mais próximos do padrão de teste. A busca pela vizinhança é feita utilizando uma medida de distância nessa procura.

O KNN foi proposto por Fukunaga e Narendra (1975), este é um classificador onde o aprendizado de um novo "objeto" é feito com base nos exemplos de treinamento (aprendizagem supervisionada). Onde pode ser expresso por:

$$d(X_i, Yi) = \sqrt[r]{\sum_{i=1}^{n} |X_i - Yi|^r}$$

$$d(X_{i'} Y_{i}) =$$

Onde d representa a distância, e $X_i \, e \, Y_i$ representam as instâncias, n é o número de atributos e r sendo a dimensão pertencente, podendo ocorrer certas alterações dependendo da métrica utilizada, Euclidiana, Manhattan ou então Minkowski.

3.2.4. Classificador: Support Vector Machine

Support Vector Machines (SVM) é um algoritmo supervisionado baseado na teoria do aprendizado estatístico (teoria Vapnik-Chervonenkis) projetado para tarefas de classificação, que têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes por meio da minimização dos erros (VAPNIK et al., 1995).

O SVM consiste em uma técnica computacional de aprendizado para problemas de reconhecimento de padrão. Introduzida por meio da teoria estatística de aprendizagem por Vapnik et al., (1995), essa classificação é baseada no princípio de separação ótima entre classes, tal que se as classes são separáveis, a solução é escolhida de forma a separar o máximo as classes.

O SVM busca um hiperplano ótimo como uma função de decisão em um espaço de características que pode ter muitas dimensões (BOSER et al., 1992; CRISTIANINI; SHAWE-TAYLOR, 2000). Para essa otimização, o SVM introduz uma minimização do risco estrutural, do inglês structural risk minimization (SRM), considerando o melhor separador, aquele que minimiza o erro de generalização e tentando evitar problemas de *overfitting* (GUO et al., 2019).

3.3. Validação de classificadores e busca de hiperparâmetros

Grande parte dos estudos que envolvem algoritmos de aprendizagem de máquina têm como objetivo estruturar modelos capazes de capturar padrões característicos de determinada base de dados para poder realizar classificação ou regressão (CLAESEN; DE MOOR, 2015). Em modelos de redução de dimensionalidade, é comum que haja propostas arquiteturais e de hiperparametrização. A arquitetura varia entre a quantidade de dimensões e classes, enquanto a hiperparametrização é expressa por elementos como taxa de aprendizagem, função de ativação, função de perda/custo, dentre outros.

3.3.1. *Grid Search*

A busca por hiperparâmetros é comumente realizada de forma manual por meio de testes dentro de um conjunto de parâmetros pré-definidos (HINTON, 2012). Uma das técnicas que automatiza esta tarefa é denominada *Grid Search*, a qual garante escolhas ótimas de um conjunto pré-estabelecido de possibilidades (BERGSTRA; BENGIO, 2012). Na prática, o que um algoritmo de busca de hiperparametrização faz é encontrar o conjunto de parametrização que minimize o erro de generalização do modelo, por meio do processo denominado como busca de hiperparâmetros ótimos (BERGSTRA; BENGIO, 2012).

Basicamente o algoritmo constrói uma grade discreta combinando alguns parâmetros e resolvendo o modelo para cada combinação de parâmetros. A escolha dos parâmetros ótimos baseia-se em qual combinação atinge o melhor desempenho

em termos de validação, ou seja, aqueles parâmetros que produzem maior acurácia global. Apesar de sua complexidade computacional elevada e de sua natureza heurística este algoritmo vem sendo amplamente utilizado devido a sua facilidade de implementação e possibilidade de ser utilizado em conjunto com diversos *solvers* (MOORE et al., 2011).

3.3.2. *Cross Validation*

A validação dos modelos de aprendizagem de máquina, referente ao conjunto de dados, dá-se pelo particionamento do conjunto de dados em três partes: conjunto de treinamento, de teste e de validação. A aplicação deste particionamento em um conjunto de dados relativamente pequeno pode levar a resultados não confiáveis (BISHOP, 2006).

Para avaliar a qualidade de um modelo gerado pelos classificadores e redutores, em termos de classificação, uma alternativa muito popular é a *Cross-Validation* (CV) (STONE, 1974). A técnica tem como característica fazer uma melhor varredura da base de dados, atenuando dessa maneira problemas causados por diferenças na base, procurando assim evitar o problema de *overfitting* na base (MOORE, 2001).

O método da validação cruzada identifica o melhor modelo para representar os dados reais, permitindo assim que se façam boas inferências. Esse método consiste em retirar um ponto da amostra e estimá-lo verificando a diferença entre o ponto amostrado e o estimado, faz-se isso com todos os pontos separadamente.

Um procedimento frequentemente utilizado para a solução desse problema é o *k-fold cross-validation (k-fold)*. O *k-fold* divide o conjunto de dados em *k* conjuntos. Assim, treina *k* modelos com os $k - 1$ conjuntos e, em cada iteração do *k-fold*, o conjunto restante é utilizado como teste. A avaliação geral é dada pela média dos *k* modelos (SHALEV-SHWARTZ; BEN-DAVID, 2013). A Figura 7 ilustra o procedimento de particionamento do *k-fold.*

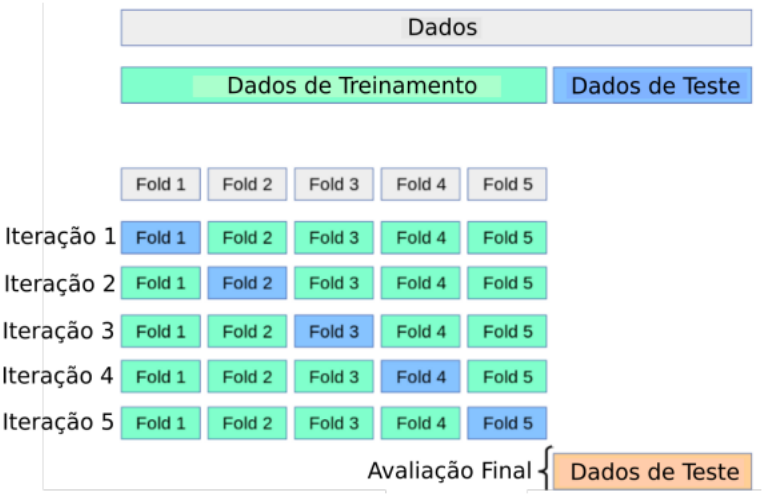

Figura 7. Processo de particionamento do conjunto de dados k-fold.

Fonte: Adaptado e traduzido de (PEDREGOSA et al., 2011).

Na Figura 7, o conjunto de dados é disposto em uma lista com todos os exemplos. Assim, o k-fold vai particioná-la em k partes formando os conjuntos de exemplos. Os $k-1$ conjuntos serão usados para treinamento do modelo e o k-ésimo restante para teste. Os componentes de $k-1$ serão mudados um de cada vez até formar os k modelos e cada modelo terá sua própria avaliação das métricas. O valor de k é frequentemente escolhido como 5, 10 ou 20 (ANGUITA et al., 2012), para o presente estudo foi escolhido *k=10*.

O processo é realizado K vezes, assim garantindo que todos os dados coletados sejam ao menos uma vez o "conjunto de testes". No final das K iterações, o Erro Médio Quadrático (*MSE - Mean Square Error*) é calculado considerando os conjuntos, o processo descrito, avalia a capacidade de predição do modelo.

3.4. Trabalhos relacionados

3.4.1. Um novo método Wrapper multiobjetivo para seleção de bandas de Imagens Hiperespectrais, por Saqui (2020)

Nesta pesquisa foi elaborado um método de seleção de bandas multiobjetivo chamado *Wrapper Multiobjective Evolutionary Band Selection* (*WMoEBS*) composto por estratégias que foram testadas experimentalmente. O *WMoEBS* é baseado na estratégia Wrapper incorporando o classificador Support Vector Machine (SVM), que utiliza informação espacial e espectral, e realiza uma seleção inicial para diminuir as bandas correlacionadas, consistindo num algoritmo multiobjetivo para lidar com

resultados da classificação e quantidade de bandas simultaneamente e um tomador de decisão para retornar uma única solução final.

3.4.2. Extração de atributos em imagens de sensoriamento remoto utilizando Independent Component Analysis e combinação de métodos lineares, por Levada (2006)

O presente trabalho de Levada, apresenta uma metodologia para melhorar o desempenho da classificação criando modelos para fusão de atributos que combinam métodos estatísticos de segunda ordem com métodos de ordens superiores, superando limitações existente nas abordagens tradicionais, como problemas de mal-condicionamento, o que pode provocar instabilidade na estimação dos componentes independentes, além de eventuais amplificações de ruídos. O esquema resultante é utilizado para combinar atributos obtidos através de diversos métodos num único vetor de padrões em duas abordagens: Fusão Concatenada e Fusão Hierárquica

3.4.3. Redução de dimensionalidade em imagens hiperespectrais usando Codificadores automáticos, por Kakarla et al., (2020)

Este artigo apresenta uma análise não linear para redução de dimensionalidade usando codificadores automáticos. O desempenho do modelo proposto é comparado com outros métodos popularmente usados como PCA e kernel PCA (KPCA) usando os classificadores KNN, KNN ponderado em conjuntos de dados de benchmark obtidos da Repositório Computational de dados Intelligence Group (CIG). Experimentalmente, foi provado que a técnica proposta usando auto-codificadores supera as técnicas de redução de dimensionalidade existentes PCA e KPCA.

3.4.4. Demais trabalhos

O emprego de IHs em estudos florestais vem crescendo à medida que aumenta a exigência de um detalhamento maior da estrutura das florestas, de modo a ser cada vez mais eficiente, sendo possível a discriminação e reconhecimento de características dos vegetais, guiando a uma análise mais precisa da composição e condições sanitárias da floresta (PETEAN, 2015).

Outro uso para IHs está voltado ao ambiente urbano, possibilitando a distinção entre os mais diversos componentes da paisagem urbana de maneira confiável (PETEAN, 2015). De acordo com Herold et al. (2003), os ambientes urbanos representam uma das mais desafiadoras áreas de análise para o sensoriamento remoto, pois sua diversidade espectral excede àquela encontrada nos ambientes naturais.

E, além disso, têm-se processos aos quais usam da espectroscopia de IHs e a quimiometria para a quantificação e classificação no ramo da agroindústria, como: identificação de fontes de licopeno e carotenóides, como o betacaroteno (BARANSKA et al., 2006), classificação de azeite extravirgem (SINELLI et al., 2010), determinação de parâmetros de qualidade em produtos lácteos (RŮŽIČKOVÁ; SUSTOVÁ, 2006), caracterização de azeitona de mesa (CASALE et al., 2010), determinação de açúcar em uva (JARÉN et al., 2001), quantificação do teor de proteína em produtos de leite em pó (INGLE et al., 2016), controle de qualidade de extratos de frutos silvestres durante o armazenamento (GEORGIEVA et al., 2014).

4. METODOLOGIA E ESTRATÉGIA DE AÇÃO

A fim de alcançar os objetivos descritos anteriormente, a metodologia foi conduzida conforme as etapas estabelecidas por Fayyad et al. (1996): Pré-processamento, Transformação, Mineração de dados, Avaliação e Interpretação (Figura 8).

Para a produção do projeto foi utilizado a ferramenta *Google Colab*, voltado para o desenvolvimento dos algoritmos feitos na linguagem Python, na versão 3.8, utilizando as bibliotecas: Pandas para a análise dos dados, Numpy para cálculo em vetores multidimensionais, Matplotlib e Seaborn para representação gráfica dos dados, e também a Sklearn para o uso dos redutores e classificadores.

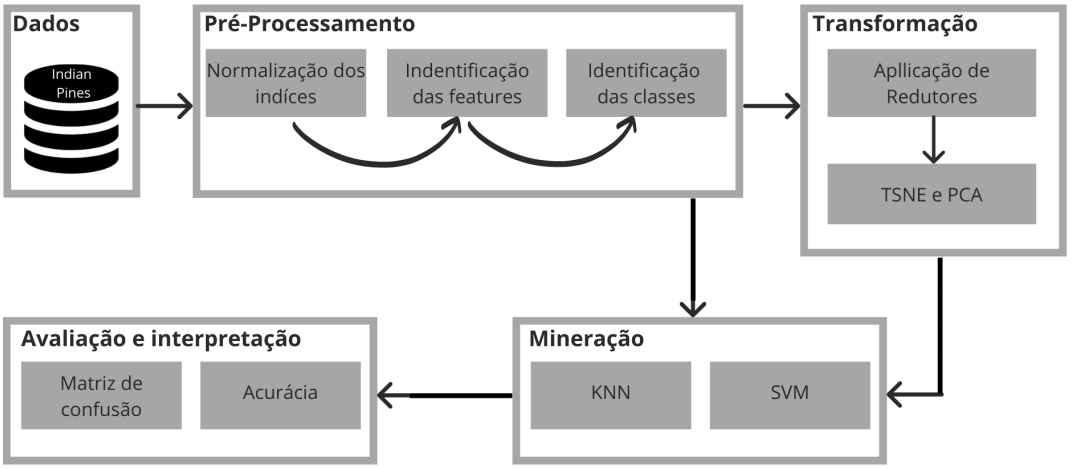

Figura 8. Fluxograma descrevendo as etapas de Análise e Descoberta de Conhecimento (KDD) para as bases de dados Indian Pines.

Fonte: Do autor.

Para Fayyad et al. (1996), KDD é o processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados. A KDD é a extração da informação interessante ou padrões dos dados em bases de dados.

4.1. Dados

A área sob estudo foi coletada em Junho de 1992 pelo sensor AVIRIS na região noroeste do estado americano de Indiana, a cena identificada como Indian Pines é composta de 145 × 145 pixels e 224 bandas de refletância com comprimento de onda no intervalo de 0,4 a 2,5 μm, sendo que a base de dados utilizada é a imagem tratada em *Comma-separated values* (CSV) com 200 bandas espectrais, esta diferença de 24 bandas se dá ao fato da remoção de bandas ruidosas as quais cobrem a região de absorção de água.

A cena de Indian Pines contém cerca de dois terços sendo agricultura e cerca de um terço sendo floresta ou e ainda uma parte sendo vegetação natural perene. Existem duas principais rodovias de pista dupla, uma linha ferroviária, bem como algumas habitações de baixa densidade, outras estruturas construídas e estradas menores. Como a cena é tirada em junho, há algumas culturas presentes, como milho, soja, que estão em estágios iniciais de crescimento com menos de 5% de cobertura. A verdade básica disponível é dividida em dezesseis classes e nem todas são mutuamente exclusivas.

Para este trabalho, cada instância (amostra) de dados para os classificadores

é um pixel da imagem Indian Pines, portanto, o total de instâncias utilizadas é de 21025 pixels. Uma visualização em tons de cinza é dada na Figura 9.a. Uma visualização em falsa composição *Red Green Blue* (RGB) é dada na Figura 9.b. Uma visualização da rotulação da região é dada na Figura 9.c. A identificação dos rótulos da IH é dada na Tabela 1.

Figura 9: Visualização mapa de Indian Pines

a) Indian Pines - imagem em tons de cinza. Extraída de (BAUMGARDNER; BIEHL; LANDGREBE, 2015)

b) Indian Pines - falsa composição RGB. (SAQUI, 2018)

c) Indian Pines - mapa de rótulos

Rótulo	Descrição	Rótulo	Descrição
1	Alfafa	9	Aveia
2	Milho - primeira fase	10	Soja - primeira fase
3	Milho - segunda fase	11	Soja - segunda fase
4	Milho - terceira fase	12	Soja - terceira fase
5	Grama - pastagem	13	Trigo
6	Grama - árvores	14	Bosques
7	Grama - pastagem cortada	15	Construções - Grama - Árvores - Ruas
8	Feno	16	Rochas - Estruturas Férreas - Edifícios

Tabela 1: Indian Pines - Rótulos das classes

4.2. Pré-Processamento

O processo de pré-processamento dos dados foi realizado pela Universidade de Purdue (noroeste do estado de Indiana, Estados Unidos) para mitigar os efeitos de detectores ruins, descalibração inter-detector e anomalias intermitentes. Bandas não calibradas e ruidosas correspondentes à absorção de água foram removidas e as 200 bandas remanescentes foram incluídas como candidatas no estudo.

O pré-processamento da cena hiperespectral abrangeu as etapas de: correção radiométrica pelo software CaliGEO Pro 2.2, conforme Campos (2017), correção atmosférica (software ATCOR-4) e correção geométrica. Além disso, aplicou-se o filtro Flat Bottom Smoother para suavizar as curvas de reflectância de cada banda, pela comparação entre três pontos (antecessor, atual e posterior), caso o ponto atual possua o menor valor de reflectância de seus vizinhos este será substituído pelo menor valor desses (SILVA, 2017), por fim foi utilizada a versão corrigida da imagem cujas bandas que cobrem regiões de absorção de água foram removidas (bandas [104-108], [149-163], 200).

4.3. Transformação

A transformação pautou-se em reduzir a dimensionalidade da IH de *Indian Pines* utilizando separadamente o PCA e o t-SNE, sendo os mesmos aplicados de forma independente.

Na aplicação do PCA inicialmente a imagem contava com 200 bandas espectrais (dimensões) e então foi reduzido para apenas 10 componentes (dimensões).

Já na redução utilizando o t-SNE foi feita originalmente a IH com 200 dimensões para apenas duas dimensões, tendo como parâmetros, 10 de perplexidade em 1200 iterações, reduzindo a base em 99% dos atributos originais.

4.4. Mineração

A mineração da imagem, divide-se em classificar os pixels sem antes reduzir a dimensionalidade, para posteriormente classificá-la reduzida, e então realizar comparativos com os dados classificados com e sem redução. Para isto fez-se uso do KNN com métrica euclidiana, analisando os pixels vizinhos de um intervalo de 100 à 1000, e obtendo a acurácia pela média da iteração, e já o SVM operou com uma taxa de regularização de 3500, utilizando um kernel radial, com coeficiente de

kernel escalar, e ocupando 16 gigabytes de cache para a execução, isso após a busca de hiperparâmetros com *Grid Search*.

Abaixo têm-se a lista de cada caso das execuções:

Caso 1: aplicação do KNN; Caso 2: aplicação do SVM;

Caso 3: aplicação do PCA; Caso 4: aplicação do t-SNE;

A hiperparametrização necessária nesses modelos foi definida por meio do Grid Search e k-fold cross-validation a fim de encontrar um conjunto de hiperparâmetros com o menor erro na função de custo. Para isso, foi considerado previamente um conjunto de hiperparâmetros a serem testados pelo Grid Search e avaliados pelo *k-fold cross-validation.*

A Figura 10 apresenta como *GridSearch* e o *k-fold cross-validation* podem ser utilizados em conjunto para otimizar um conjunto de hiperparâmetros. Primeiro são definidos os conjuntos de hiperparâmetros iniciais. Após, é calculado o produto cartesiano desses conjuntos e, para cada elemento do produto cartesiano, o erro é estimado pelo *k-fold* cross-validation. Após todos os hiperparâmetros serem avaliados é retornado um conjunto de hiperparâmetros com o menor custo.

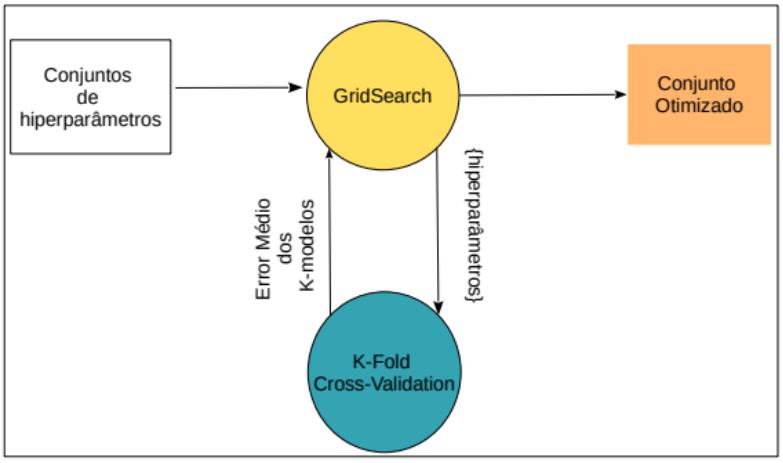

Figura 10. GridSearch e k-fold cross-validation.

Fonte: Do autor

4.4.1. Escolha dos hiperparâmetros com grid search

4.4.1.1. Caso 1: KNN

Alguns dos hiperparâmetros mais comuns do KNN são: *n_neighbors* (número de vizinhos), *weights* (função de peso usada na previsão) que pode ser definido como *"uniform"*, onde cada vizinho dentro da fronteira carrega o mesmo peso ou "*distance*", onde pontos mais próximos terão maior peso para a decisão, e por fim o último hiperparâmetro '*algorithm*' definido como '*auto*', que tentará decidir qual

algoritmo (*ball_tree, kd_tree, e brute*) mais apropriado com base nos valores passados para o método de *fit* (função para ajuste do modelo). A tabela abaixo demonstra de forma sucinta os parâmetros escolhidos e os intervalos dos mesmos:

Hiperparâmetros	**Conjunto**
Número de vizinhos	[100, 1000]
Função de peso	{'*uniform*', 'distance'}
Algoritmo	{'auto'}

Tabela 2: Hiperparâmetros x Conjunto - Caso 1

4.4.1.2. Caso 2: SVM

Alguns dos hiperparâmetros mais comuns do SVM são: C (taxa de regularização), Kernel que pode ser definido como *'linear', 'poli', 'rbf', 'sigmoid', 'pré-computado'*, Gamma (coeficiente para o kernel usado), e por fim o tamanho da cache de memória utilizada. A tabela abaixo demonstra de forma sucinta os parâmetros escolhidos e os intervalos dos mesmos:

Hiperparâmetros	**Conjunto**
Taxa de regularização	{ 500, 1000, 1500, 2000, 2500, 3000, 3500 }
Kernel	{ 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' }
Gamma	{auto, 0.5, 1}
Tamanho de cache	65536

Tabela 3: Hiperparâmetros x Conjunto - Caso 2

4.4.1.3. Caso 3: PCA

Um dos hiperparâmetros mais comumente utilizados no PCA é a quantidade de componentes, que no caso representa para quantas dimensões o conjunto de dados será reduzido. A tabela abaixo demonstra de forma sucinta o parâmetro escolhido e os intervalos dos mesmos:

Hiperparâmetros	Conjunto
Número de componentes	{ 5, 10, 50, 100, 150 }

Tabela 4: Hiperparâmetros x Conjunto - Caso 3

4.4.1.4. Caso 4: t-SNE

Alguns dos hiperparâmetros mais utilizados no t-SNE são: *número de componentes*, pode receber valor dois ou três e representa assim como o PCA, a quantidade de dimensões para que o conjunto de dados será reduzido, *número de iterações*, que representa quantas iterações serão necessárias para a otimização, perplexidade está relacionada ao número de vizinhos mais próximos que é usado em outros algoritmos de aprendizagem múltiplos (recomenda-se utilizar valores entre 5 e 50), e por fim um valor de estado aleatório, que gera valores aleatórios para a inicialização.

Hiperparâmetros	Conjunto
Número de componentes	{ 2, 3 }
Número de interações	{10, 50, 100, 500, 1000}
Perplexidade	{ 5, 10, 25, 50}
Estado aleatório	{ 10, 20, 50, 100 }

Tabela 5: Hiperparâmetros x Conjunto - Caso 4

4.4.2. Validação dos resultados com cross validation

Para todos os métodos e componentes avaliados neste trabalho foram utilizadas estratégias de validação cruzada com 10 grupos (*10-fold cross-validation*). Uma representação da estratégia *10-fold cross-validation* pode ser visualizada na Figura 11, onde para cada *fold* (grupo de pixels), os pixels de amostras separados para treinamento formaram um montante de 90%, ou seja, $9/10$ da quantidade total de pixels. Esses pixels são representados pelos blocos azuis da Figura 11 e foram utilizados durante cada um dos processos de seleção de bandas, sendo divididos em dois subgrupos, sendo um para treinamento e outro para validação.

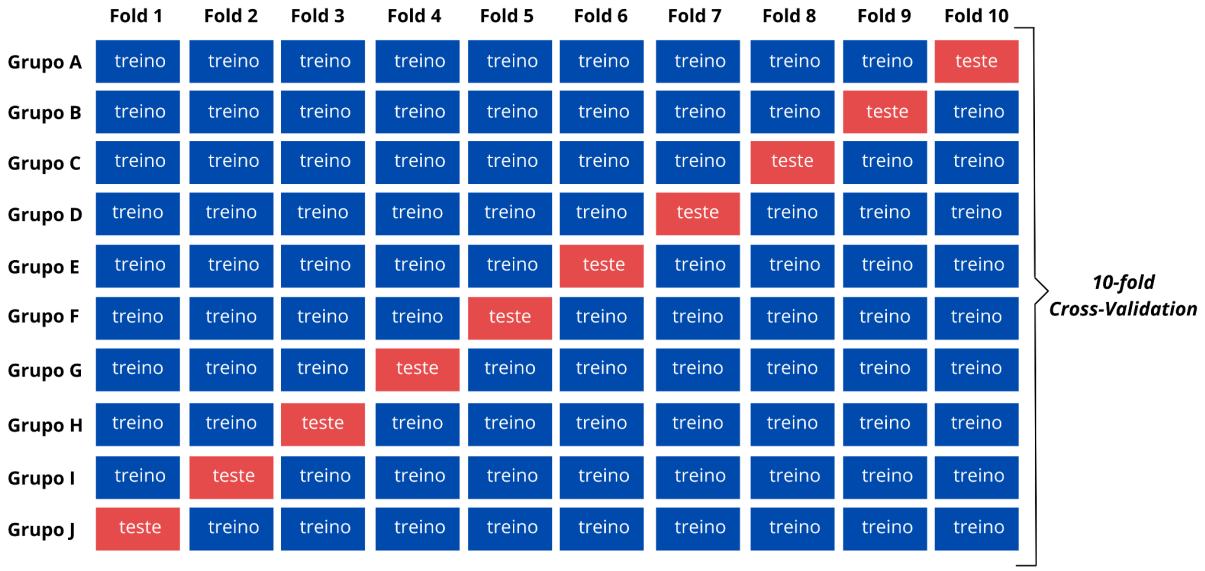

Figura 11. Configuração das amostras para os experimentos com 10-fold Cross-Validation

Fonte: do Autor

Os pixels de teste estão representados pela cor vermelha na Figura 11 e uma característica importante, foi a forma como os 10%, ou seja, 1/10 dos pixels de testes de cada um dos folds foram organizados. Para essa organização, uma estratégia de separação foi aplicada por meio de amostragem estratificada e sistemática para formar cada um dos *folds*, sendo essa estratégia descrita na sequência:

- Primeiramente, os pixels (amostras) foram organizados em classes obtidas a partir do GT;
- Então, em cada classe, os pixels foram ordenados conforme as posições originais na imagem, observando primeiro as colunas e depois as linhas;
- Por fim, para cada uma das classes, os pixels eram selecionados alternadamente para compor cada um dos 10 folds.

Esse procedimento foi repetido até que cada fold contivesse 10% dos pixels de testes, ou seja, por meio dessa organização cada fold era totalmente diferente dos demais.

Para cada *fold*, após a seleção de bandas, os 90% dos pixels (amostras de treinamento) são novamente utilizados para o treinamento de um modelo classificador. Nesse momento é utilizado o mesmo classificador do processo de

seleção de bandas, porém agora treinado com as bandas selecionadas e usando de uma única vez todos os 90% dos pixels. Após essa realização, os outros 10% do total dos pixels (amostras de testes) são utilizados para os testes do conjunto de bandas selecionados e avaliação do desempenho de classificação para cada um dos métodos e casos considerados.

Essas configurações de distribuições dos pixels para cada um dos *folds* apresentadas neste subcapítulo foram obtidas uma única vez e utilizadas nos diferentes casos de seleção de bandas aplicados nos experimentos iniciais proporcionando experimentos justos. Após a realização dos testes em cada conjunto de bandas selecionado, para cada caso, a média dos diferentes índices obtidos pelo valor dos 10 grupos foi apresentado nos resultados.

4.5. Avaliação e Interpretação

A avaliação e interpretação foi realizada através da geração de matriz de confusão para compreendimento da classificação, e verificando a acurácia obtida, por meio das funções de Metrics da biblioteca Sklearn, e ainda utilizando a função de relatório de classificação, será possível também verificar quantos pixels fora classificado para cada uma das 16 classes, e quantos pertencem às mesmas. Deste modo será possível analisar o comportamento do classificador para cada resultado de classe e pixel, verificando assim para qual está mais assertivo.

6.RESULTADOS E DISCUSSÃO

Um dos aspectos em que ainda há um caminho longo para evolução na aprendizagem de máquina é em relação ao tratamento de dados com alta dimensionalidade, nesse sentido, o trabalho apresentou os redutores de dimensionalidade, t-SNE e PCA, com os classificadores KNN e SVM aplicados a imagem hiperespectral Indian Pines.

6.1. Resultados Alcançados

Abaixo apresenta-se a tabela da média da acurácia dos 10 *k-fold* gerado no processo de *Cross Validation* com os hiperparametros escolhidos com *Grid Search* da classificação de Indian Pines (esforço computacional) e ainda acurácia esperada foi obtida pela por meio dos trabalhos relacionados do referencial teórico, utilizando os classificadores KNN e SVM, com e sem os redutores de dimensionalidade, PCA e

t-SNE.

	Obtido	Esperado
Algoritmo	**Acurácia da Classificação**	
KNN	72,58%	≥ 78,84% (HUANG et al., 2016)
SVM	88,49%	≥ 94,71% (HUANG et al., 2016)
PCA + KNN	70,07%	≥ 71,83% (ABOAGYE et al., 2012)
PCA + SVM	77,67%	≥ 79,02% (CHEN, 2021)
t-SNE + KNN	73,23%	≥ 71,04% (HARIHARAN et al., 2021)
t-SNE + SVM	92,85%	≥ 93,35% (GAO et al., 2019)

Tabela 6: Média da acurácia da classificação de Indian Pines

Evidentemente o classificador SVM obteve um resultado proeminente mais satisfatório do que em relação aos demais, em questão de o mesmo têm uma maior gama de hiperparâmetros para a classificação, necessitando assim de otimização dos hiperparâmetros para o caso, o que já não ocorre ao KNN, pois o mesmo leva em conta apenas a distância de seus vizinhos para a classificação, que no estudo fora estimado da média seus seus primeiros 1000 vizinhos (ao qual representa pouco menos de 5% da amostra total), tendo carência na precisão nos resultados colhidos em questão da redução de dimensionalidade, ao qual foi utilizado Grid Search para a escolha dos hiperparâmetros, juntamente com Cross Validation para a validação cruzada dos resultados em ambos os casos, conseguindo assim mais de 90% de acurácia na classificação de Indian Pines, reduzida com t-SNE e classificada com SVM.

Portanto, a partir dos resultados obtidos evidencia o fato do uso de RD para o caso de Indian Pines em específico, de modo a obter uma classificação medianamente precisa, e para a otimização da classificação é explícito que o uso do algoritmo de manifold learning, t-SNE melhora a acurácia.

7.CONSIDERAÇÕES FINAIS

O trabalho buscou avaliar a redução da dimensionalidade dos dados em imagens hiperespectrais (IHs) a partir do uso de algoritmos baseados em Manifold

Learning para o propósito de classificação de IHs, com isso introduziu formalmente o problema de RD e sua importância na análise de IHs. Uma técnica da classe de métodos de Manifold Learning é aplicada, t-SNE, e também um segundo método popularmente conhecido no campo de RD, o PCA. De modo a evidenciar para a comunidade que necessita de RD (pesquisadores e cientistas estatísticos, cientistas de *Machine Learning,* desenvolvedores que precisam de ML, e etc…) o comportamento dos algoritmos empregados.

Como já foi dito, a metodologia proposta é dirigida ao processo de classificação de dados em alta dimensionalidade (imagens hiperespectrais). Como é bem conhecido, o atrativo da utilização destes dados reside em possibilitar a separação de classes com características espectrais muito semelhantes (FUKUNAGA, 1990), não separáveis quando se emprega dados convencionais em baixa ou média dimensionalidade. Esta vantagem fica, entretanto, prejudicada em classificadores paramétricos (caso do classificador KNN) sempre que o número de amostras de treinamento disponíveis não é limitado, fato este que ocorre com frequência em situações reais, resultando no conhecido problema conhecido por fenômeno de Hughes. É no fato de que tal fenômeno não atinge classificadores não-paramétricos que reside a vantagem do SVM.

Estudos experimentais foram conduzidos em conjuntos de dados de IHs amplamente usados, Indian Pines, que são retirados do Grupo de Inteligência Computacional (CIG), coletadas pelo sensor AVIRIS na região noroeste do estado americano de Indiana, para o estudo da RD foram usados os redutores PCA e t-SNE juntamente com os classificadores KNN e SVM, os resultados são comparados com técnicas convencionais de redução de dimensionalidade, como é o caso do PCA, em conjunto ao método de Manifold Learning, o t-SNE.

Os resultados empíricos mostram que o t-SNE tem um bom desempenho em comparação com o PCA, tendo seus hiperparâmetros escolhidos através do uso de Grid Search e o resultados validados por meio da média dos *k-folds* na *Cross Validation,* projetando dados em um novo espaço de recursos usando diferentes kernels, provendo uma melhor acurácia, e com a utilização do redutor de dimensionalidade t-SNE com o classificador SVM obtendo uma acurácia de 92,85%, enquanto o PCA com KNN tendo cerca de 70,07% de acurácia. A técnica proposta pode ser facilmente implementada e funciona de forma eficiente para redução de dimensionalidade.

8. REFERÊNCIAS BIBLIOGRÁFICAS

ABOAGYE, Samuel Opoku; CUI, Suxia. **Hyperspectral Image Feature Extraction and Selection Using Empirical Mode Decomposition PCA.** In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012. p. 1.

ANGUITA, D. et al. **The 'K'in K-fold cross validation**. In: ESANN, 2012, Bruges.

BARANSKA, Malgorzata; SCHÜTZE, W.; SCHULZ, Hartwig. **Determination of lycopene and β-carotene content in tomato fruits and related products: comparison of FT-Raman, ATR-IR, and NIR spectroscopy.** Analytical Chemistry, v. 78, n. 24, p. 8456-8461, 2006.

BAUMGARDNER, Marion F.; BIEHL, Larry L.; LANDGREBE, David A. **220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3**. Purdue University Research Repository, v. 10, p. R7RX991C, 2015.

BELLMAN, Richard E. **Adaptive control processes: a guided tour.** Princeton university press, 2015.

BERGSTRA, James; BENGIO, Yoshua. **Random search for hyper-parameter optimization.** Journal of machine learning research, 2012, 13.2.

BISHOP, C. M. Patter **Recognition and Machine Learning.** New York: Springer, 2006.

BORGES, HELYANE BRONOSKI; NIEVOLA, J. C. Redução de Dimensionalidade em Bases de Dados de Expressão Gênica. 2006. Tese de Doutorado. Dissertação de Mestrado, PPGIa-PUCPR.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N. **A training algorithm for optimal margin classifiers.** In: Proceedings of the fifth annual workshop on Computational learning theory. 1992. p. 144-152.

CAMPOS, T. L de L. **Discriminação de espécies arbóreas nativas da Floresta Estacional Semidecidual e da exótica Eucalyptus urograndis W. Hill ex Maiden utilizando dados hiperespectrais**. Dissertação (Mestrado em Agronomia). Universidade Estadual de Maringá (UEM), Maringá PR, 121f, 2017.

CASALE, Monica et al. **Characterisation of table olive cultivar by NIR spectroscopy.** Food chemistry, v. 122, n. 4, p. 1261-1265, 2010.

CAYTON, Lawrence. **Algorithms for manifold learning.** Univ. of California at San Diego Tech. Rep, v. 12, n. 1-17, p. 1, 2005.

CHEN, Guang Yi. **Multiscale filter-based hyperspectral image classification with PCA and SVM.** Journal of Electrical Engineering, v. 72, n. 1, p. 40-45, 2021.

CLARK, Roger N. et al. **Spectroscopy of rocks and minerals, and principles of spectroscopy.** Manual of remote sensing, v. 3, n. 3-58, p. 2-2, 1999.

CLAESEN, Marc; DE MOOR, Bart. **Hyperparameter search in machine learning**. arXiv preprint arXiv:1502.02127, 2015.

CRISTIANINI, Nello; SHAWE-TAYLOR, John. **Support Vector Machines and other kernel-based learning methods**. Cambridge, 2004.

DINIZ, Fábio Abrantes et al. **RedFace: um sistema de reconhecimento facial baseado em técnicas de análise de componentes principais e autofaces.** Revista Brasileira de Computação Aplicada, v. 5, n. 1, p. 42-54, 2013.

DISNER, Daniel da Silva. **Mineração de dados para obtenção de conhecimento em Big Data.** 2015.

DONOHO, David L.; GRIMES, Carrie. **Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences**, v. 100, n. 10, p. 5591-5596, 2003.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases.** AI magazine, v. 17, n. 3, p. 37-37, 1996.

FLORENZANO, Teresa Gallotti. **Iniciação em sensoriamento remoto.** Oficina de textos, 2007.

FUKUNAGA, Keinosuke; NARENDRA, Patrenahalli M.. . **A branch and bound algorithm for computing k-nearest neighbors.** IEEE transactions on computers, v. 100, n. 7, p. 750-753, 1975.

FUKUNAGA, Keinosuke. **Introduction to statistical pattern recognition.** Elsevier, 2013.

GAO, Yanlong; FENG, Yan; YU, Xumin. **Feature extraction and classification of hyperspectral images using hierarchical network.** IEEE Geoscience and Remote Sensing Letters, v. 17, n. 2, p. 287-291, 2019.

GEORGIEVA, Mariya et al. **Application of NIR spectroscopy and chemometrics in quality control of wild berry fruit extracts during storage.** Hrvatski časopis za prehrambenu tehnologiju, biotehnologiju i nutricionizam, v. 8, n. 3-4, p. 67-73, 2013.

GHAMISI, Pedram et al. **Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art**. IEEE Geoscience and Remote Sensing Magazine, v. 5, n. 4, p. 37-78, 2017.

GUO, Yanhui et al. **Hyperspectral image classification with SVM and guided filter.** EURASIP Journal on Wireless Communications and Networking, v. 2019, n. 1, p. 1-9, 2019.

HARIHARAN, S. et al. **Analysing Effect of t-SNE and 1-D CNN on Performance of Hyperspectral Image Classification.** Turkish Journal of Computer and Mathematics Education (TURCOMAT), v. 12, n. 6, p. 1828-1833, 2021.

HEROLD, Martin et al. **The spectral dimension in urban land cover mapping from high-resolution optical remote sensing data**. In: Proceedings of the 3rd Symposium on remote Sensing of Urban Areas. 2002. p. 2002.

HINTON, Geoffrey E. **A practical guide to training restricted Boltzmann machines.** In: Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 2012. p. 599-619.

HONGYU, Kuang. **Comparação do GGE-biplot ponderado e AMMI-ponderado com outros modelos de interação genótipo × ambiente.** 2015. Tese de Doutorado. Tese (Doutorado em Estatística e Experimentação Agronômica). Piracicaba: USP. 155p.

HOTELLING, Harold. **Analysis of a complex of statistical variables into principal components.** Journal of educational psychology, v. 24, n. 6, p. 417, 1933.

HOTELLING, Harold. **Simplifield calculation of principal components**. Psychometrika, Williamsburg, v.1, p.27-35, 1936.

HUANG, Kunshan et al. **Spectral–spatial hyperspectral image classification based on KNN**. Sensing and Imaging, v. 17, n. 1, p. 1-13, 2016.

HUGHES, Gordon. **On the mean accuracy of statistical pattern recognizers.** IEEE transactions on information theory, v. 14, n. 1, p. 55-63, 1968.

INFO ESCOLA. **Espectro Eletromagnético**. Disponível em: https://www.infoescola.com/fisica/espectro-eletromagnetico/. Acesso em: 24 nov. 2021.

INGLE, Prashant D. et al. **Determination of protein content by NIR spectroscopy in protein powder mix products.** Journal of AOAC International, v. 99, n. 2, p. 360-363, 2016.

JACOBS, Adam. **The pathologies of big data**. Communications of the ACM, v. 52, n. 8, p. 36-44, 2009.

JARÉN, C. et al. **Sugar determination in grapes using NIR technology.** International Journal of Infrared and Millimeter Waves, v. 22, n. 10, p. 1521-1530, 2001.

JIMENEZ, Luis O.; LANDGREBE, David A. **Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data.** IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), v. 28, n. 1, p. 39-54, 1998.

JOLLIFFE, Ian T. **Principal component analysis.** Technometrics, v. 45, n. 3, p. 276, 2003.

KAKARLA, Syam et al. **Dimensionality Reduction in Hyperspectral Images Using Auto-encoders.** In: International Conference on Advances in Computational Intelligence and Informatics. Springer, Singapore, 2019. p. 101-107.

KENDALL, M. G. **A course in the geometry of n-dimensions**, N. York, Ed. 1961.

LAROSE, Daniel T. **Data mining methods and models.** Hoboken: Wiley-Interscience, 2006.

LAZER, David et al. **The parable of Google Flu: traps in big data analysis**. Science, v. 343, n. 6176, p. 1203-1205, 2014.

LEVADA, Alexandre Luís Magalhães. **Extração de atributos em imagens de sensoriamento remoto utilizando Independent Component Analysis e combinação de métodos lineares.** 2006.

MANLY, Bryan F. J.. **Multivariate statistical methods.** New York, Chapman and Hall, 1986.159 p.

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados.** Editora Casa do Código, 2016.

MOORE, Andrew W. **Cross-validation for detecting and preventing overfitting.** School of Computer Science Carneigie Mellon University, 2001.

MOORE, Gregory; BERGERON, Charles; BENNETT, Kristin P. **Model selection for primal SVM.** Machine learning, 2011, 85.1: 175-208.

MUNDOGEO. **Sensoriamento Remoto Hiperespectral.** Disponível em: https://mundogeo.com/2004/08/23/sensoriamento-remoto-hiperespectral/. Acesso em 24 jun. 2021.

PEARSON, Karl. LIII. **On lines and planes of closest fit to systems of points in space.** The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, v. 2, n. 11, p. 559-572, 1901.

PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011

PETEAN, Felipe Coelho de Souza. **Uso de imagens hiperespectrais e da tecnologia LiDAR na identificação de espécies florestais em ambiente urbano na cidade de Belo Horizonte, Minas Gerais**. 2015. Tese de Doutorado. Universidade de São Paulo.

ROSSANT, C. 2015. Disponível em: <https://www.oreilly.com/content/an-illustrated-introduction-to-the-t-sne-algorithm/>.

RŮŽIČKOVÁ, J. A. N. A.; ŠUSTOVÁ, KVĚTOSLAVA. **Determination of selected parameters of quality of the dairy products by NIR spectroscopy.** Czech journal of food sciences, v. 24, p. 255-260, 2006.

SANDANIELO, Vera Lúcia Martins. **"Emprego de técnicas estatísticas na construção de índices de desenvolvimento sustentável aplicados a assentamentos rurais."** (2008): xv-159.

SAQUI, Diego. **Metodologia supervisionada para seleção de bandas de imagens hiperespectrais utilizando o NSGA2.** Tese (Doutorado) - Universidade Federal de São Carlos, January 2018.

SAUL, Lawrence K.; ROWEIS, Sam T. **Think globally, fit locally: unsupervised learning of low dimensional manifolds.** Departmental Papers (CIS), p. 12, 2003.

SCOTT, David W. **Multivariate density estimation and visualization.** In: Handbook of computational statistics. Springer, Berlin, Heidelberg, 2012. p. 549-569.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning: From theory to algorithms.** 2013.

SILVA, L. C. D. A. **Análise e categorização dos padrões espectro-temporais de índices de vegetação de alvos agrícolas e permanentes oriundos do sensor Modis entre os anos-safras 2013/2014 e 2016/2017.** Trabalho de conclusão de curso (Engenharia agrícola) Universidade do oeste do paraná (UNIOESTE), Cascavel PR, 128p, 2017.

SINELLI, Nicoletta et al. **Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity.** Food research international, v. 43, n. 1, p. 369-375, 2010.

STONE, M. **Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion).** Journal of the Royal Statistical Society: Series B (Methodological), 1976, 38.1: 102-102.

VAN DER MAATEN, Laurens; HINTON, Geoffrey. **Visualizing data using t-SNE.** Journal of machine learning research, v. 9, n. 11, 2008.

VAPNIK, Vladimir; GUYON, Isabel; HASTIE, Trevor. **Support vector machines**. Mach. Learn, v. 20, n. 3, p. 273-297, 1995.

ZHANG, Bing et al. **Hyperspectral image processing and analysis system (HIPAS) and its applications.** Photogrammetric engineering and remote sensing, v. 66, n. 5, p. 605-610, 2000.

Muzambinho, 09 de Janeiro de 2022