10.18605/2175-7275/cereus.v13n3p283-301

<< Recebido em: 23/09/2021 Aceito em: 01/10/2021. >>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

Aplicando Mineração de Imagens na Agricultura de Precisão

Applying Image Mining in Precision Agriculture

Diego de Castro Rodrigues[1], Marcelo Lisboa Rocha[2], Daniela Mascarenhas de Queiroz Trevisan[3], Lúcio André de Castro Jorge[4], Ednaldo José Ferreira[5], Lucas Prado Osco[6], Rommel Melgaço Barbosa[7]

RESUMO

Análise de imagens de plantações estão consolidados no mercado da agricultura de precisão. Nesse sentido, a utilização de técnicas de processamento de imagem, mineração de imagem e inteligência artificial são ferramentas fundamentais. Podendo aplicar essas técnicas de maneira individual ou em conjunto. Um problema comum em análises de imagens é que pequenas mudanças na iluminação e no momento de tirar fotos podem influenciar como as técnicas computacionais identificam seus elementos. O custo é muito alto ou mesmo inviável para identificar ou segmentar uma imagem de forma universal. Sendo assim, é necessário um ponto de partida solido para guiar as técnicas existentes. Este estudo apresenta um experimento utilizando técnicas de mineração de imagens, associado a algoritmos de associação customizado. Utilizando o conhecimento do especialista para criar e rotular conjunto de pixel de interesse. Assim, ao processar uma imagem as classes de interesse são facilmente identificadas e ajustadas para cada realidade. Os resultados empíricos indicam que nossa solução aprimora a forma de seleção de padrões identificando as classes de interesse, identificando de maneira correta solo e vegetação. Os testes foram realizados em sete mosaicos diferentes da mesma plantação. O processo de identificação das classes desejadas (solo, plantação), ocorreram de maneira satisfatória validado assim nosso estudo como uma solução viável para agricultura de precisão.

Palavras-chave: Mineração de Dados. Mineração de Imagem. Agricultura de Precisão. Regras de Associação.

ABSTRACT

Crop image analysis are consolidated in the precision farming market. In this sense, the use of image processing techniques, image mining and artificial intelligence are fundamental tools. Being able to apply these techniques individually or together. A common problem in image analysis is that small changes in lighting and timing can influence how computational techniques identify its elements. The cost is too high or even unfeasible to universally identify or segment an image. As such, a solid starting point is needed to guide existing techniques. This study presents an experiment using image mining techniques, associated with custom association algorithms. Using expert knowledge to create and label pixel set of interest. Thus, when processing an image, the classes of interest are easily identified and adjusted for each reality. The empirical results indicate that our solution improves the way of selecting patterns by identifying the classes of interest, correctly identifying soil and vegetation. Tests were performed on seven different mosaics from the same culture. The process of identifying the desired classes (soil, plantation) occurred satisfactorily, thus validating our study as a viable solution for precision agriculture.

Keywords: Data Mining. Image Mining. Precision agriculture. Association Rules.

[1] Mestre em Modelagem Computacional de Sistemas-UFT; Doutorando em Ciência da Computação-UFG.
diego.rodrigues@ifto.edu.br
https://orcid.org/0000-0001-8396-1947

[2] Doutor em Engenharia Elétrica-COPPE/UFRJ. Programa de Pós-Graduação em Modelagem Computacional de Sistemas/UFT.
mlisboa@uft.edu.br
https://orcid.org/0000-0002-4034-0021

[3] Mestre e Doutoranda em Modelagem Computacional de Sistemas-UFT.
danielatrevisan@uft.edu.br
https://orcid.org/0000-0003-3677-0851

[4] Doutor e Pesquisador na Embrapa Instrumentação, Especialista em Inteligência artificial e Drones para Agricultura.
https://orcid.org/0000-0001-8341-3203

[5] Doutor e Pesquisador na Embrapa Instrumentação, Especialista em Inteligência artificial.
https://orcid.org/0000-0003-0277-669X

[6] Doutor e Pesquisador na Unoeste-SP, Especialista em processamento de imagens.
https://orcid.org/0000-0002-0258-536X

[7] Doutor e Pesquisador na UFG, áreas de atuação Mineração de Dados e Matemática Computacional.
https://orcid.org/0000-0002-2638-7026

1. INTRODUÇÃO

A mineração de imagens tem a tarefa de extração de conhecimento oculto, relacionando aos dados de imagens ou outros padrões não armazenados explicitamente em imagens e utiliza ideias de processamento de imagens, mineração de dados, aprendizado de máquina, bancos de dados, recuperação de imagens e visão computacional. A análise de dados de imagens tem unido conhecimento de diferentes domínios de conhecimento a tecnologias de mineração de dados (MUNAWAR et al., 2019; KUMAR; KAUR; GUPTA, 2020; KHAN; SOOMRO; ALAM, 2020; ZHANG; NIU, 2019). Estudos indicam que o desafio fundamental na mineração de imagens é determinar como a representação de pixel de baixo nível contida em uma imagem ou conjunto de imagens pode ser processada de maneira eficaz e eficiente para identificar objetos espaciais e relacionamentos de alto nível (GUREVICH; YASHINA, 2017; KAPLUN et al., 2017).

No entanto, com o crescimento contínuo do volume de informações e técnicas de análise (*Machine learning*, *Data Mining*, Visão Computacional e *Image Mining*), surgiu um desequilíbrio entre a demanda de dados, capacidade de análise e descoberta de padrões para contextos diversos. Portanto, as técnicas de análise de dados estão se tornando mais genéricas em uma tentativa de analisar qualquer tipo de dado. Levando a` insatisfação na descoberta de padrões específicos e a custos mais altos de tempo e conhecimento para identificar padrões. O fator crítico é geralmente o uso de técnicas genéricas para identificar padrões, em vez de utilizar soluções especializadas para o tipo de dado ou imagem. Para enfrentar esses desafios, os especialistas em domínios de dados como agricultura, medicina e visão computacional precisam de ferramentas especializadas para identificação de padrões em imagens e dado relacionados a sua área de interesse.

Neste artigo, apresentamos um estudo conduzido a dados e experimentos em conjunto de imagens reais de plantações. Desenvolvemos um Framework e um algoritmo de análise associativa aplicado ao contexto de identificação de áreas plantadas. Esse algoritmo identifica padrões utilizando métricas probabilísticas, poda, filtros e métricas customizadas. Que alinhadas ao nosso Framework aplica os padrões separando o solo da plantação.

As regras de associação demonstram simplicidade para compreender os padrões obtidos em campos como engenharia (XU et al., 2018), sistemas de recomendação (OSADCHIY et al., 2019), diagnósticos clínicos (LAKSHMI; VADIVU, 2017; DESHMUKH; BHOSLE, 2016a). A descoberta de padrões associativos é uma das principais tarefas da

mineração de dados e *Image Mining*, com destaque ao uso de regras de associação baseadas no modelo Apriori (HAN; KAMBER; PEI, 2012; DEY NILANJAN A et al., 2015; ZAHRADNIKOVA; DUCHOVICOVA; SCHREIBER, 2015; DESHMUKH; BHOSLE, 2016b; FATMA; NASHIPUDIMATH, 2011).

Buxton et. al (2019) apresenta algumas das limitações apontadas nos algoritmos clássicos de regras associativas, inicialmente proposto por Agrawal (AGRAWAL; SRIKANT et al., 1994). As limitações dos algoritmos tradicionais apresentam insuficiências relacionadas ao quantitativo de padrões gerados, redundâncias, seleção de melhores regras e eliminação de padrões.

Os algoritmos clássicos de regras associativa trabalham de forma simétrica. Assim as relações $X \rightarrow Y$ e $Y \rightarrow X$ podem ter o mesmo significado. X é um antecedente ou conjunto de características e Y é o consequente. Em termos práticos X poderia ser características de uma imagem (dados, meta-dados, valor pixel) e Y a classe desejada (classificação do pixel, identificação de plantas ou objetos). Desta forma um conjunto de características ou dados podem definir os padrões de uma imagem ($pixel = 30,32 , \dots ,40,35 \rightarrow arvore$), mas uma implicação inversa não seria necessariamente verdadeira por exemplo ($arvore \rightarrow pixel = 30,32$) nem todo pixel com valores $30{,}32$ vai representar uma árvore.

Contudo, o objetivo deste estudo é aprimorar os padrões associativos assimétricos em imagens aéreas relacionados ao contexto agrícola de identificação de cobertura vegetal e solo, utilizando um Framework que valoriza o conhecimento dos especialistas agrícolas. Selecionando os padrões associativos baseado em um conjunto de métricas assimétricas probabilísticas. Assim obtendo melhores resultados na identificação de padrões. Alguns destaques do estudo:

- Framework de Image Mining.
- Identificação de padrões em imagens com técnicas associativas.
- Utilização de métricas alternativas ao modelo Support/Confidence.
- Aplicação em conjunto de imagens reais.
- Classificação de Cobertura vegetal e solo em imagens aéreas.

O restante deste artigo está organizado conforme descrito a seguir. A seção 2 apresenta os conceitos básicos a respeito de regras de associação e sobre medidas de interesse objetivas e subjetivas sobre as mesmas. A seção 3 descreve a metodologia adotada nesse trabalho, composta de um Framework com 5 fases. Na seção 4 são apresentados os resultados experimentais e discussões sobre os mesmos. Finalmente na seção 5 são apresentadas as considerações finais e possíveis trabalhos futuros.

2. CONCEITOS BÁSICOS

O processo de mineração de dados em imagens é apresentado na Figura 1. As imagens de um acervo (banco de imagens) são recuperadas segundo critérios inerentes à aplicação. A seguir, uma fase de pré-processamento aumenta a qualidade dos dados, os quais são então submetidos a uma série de transformações e de extração de características que geram importantes informações a respeito das imagens. A partir destas informações, a mineração pode ser realizada através de técnicas específicas, com o intuito de descobrir padrões significativos. Os padrões resultantes são então interpretados e avaliados para a obtenção do conhecimento final, que pode ser aplicado no entendimento de problemas, na tomada de decisões ou em outras atividades estratégicas (HSU; LEE; ZHANG, 2002).

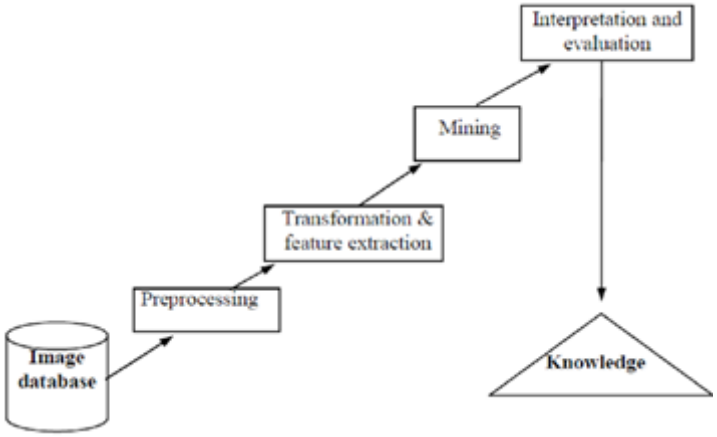

Figura 1. O processo de mineração de imagens extraído (HSU; LEE; ZHANG, 2002).

Embora o esquema da figura assemelhe-se bastante ao processo *de Knowledge Discovery in Databases* (KDD) (ASRIN et al., 2020), deve-se salientar que mineração de dados em imagens não consiste simplesmente na aplicação de técnicas de mineração de dados em bancos convencionais ao domínio de imagens. Diferenças importantes entre estes bancos e os bancos de imagens incluem, valores relativos e valores absolutos, informação espacial, Interpretação única e interpretação múltipla, Representação visual dos padrões descobertos (HSU; LEE; ZHANG, 2002).

As técnicas utilizadas pelos mineradores de imagens incluem reconhecimento de padrões, indexação e recuperação de imagens, classificação de imagem, agrupamento de imagens, mineração de regras de associação e rede neural. As técnicas são classificadas em cinco níveis de informação e nas operações associadas de mineração de imagens ou dados. Esses níveis são (nível de extração de conhecimento, padrões e nível de relações

entre imagens, nível de conceito semântico, região, objetos ou padrões visuais nível, nível de pixel).

2.1 Conceitos básicos Regras de Associação

A análise de associação é útil para descobrir relacionamentos interessantes e ocultos em grandes conjuntos de dados. Os relacionamentos descobertos podem ser representados na forma conjuntos de itens frequentes ou regras de associação. Os principais algoritmos encontrados nos estudos base para essa pesquisa são Apriori (AGRAWAL; IMIELINSKI; SWAMI, 1993) e FP-Growth (HAN et al., 2004). Adotam uma estratégia de dividir a descoberta de padrões de associação em duas etapas principais (TAN; STEINBACH; KUMAR, 2005), Etapa 1, geração dos *itemsets* frequentes e Etapa 2, geração das regras de associação.

Uma regra de associação é expressa por $X \rightarrow Y$ e X é o antecedente da regra e X é o seu consequente. O funcionamento geral dos algoritmos de regras de associação requer um valor mínimo de Suporte (Equação 1) e um valor mínimo de Confiança (Equação 2). Com o valor mínimo de Suporte é realizado o cálculo para geração de itens frequentes (Etapa 1), que utiliza o valor do Suporte mínimo como parâmetro de corte para selecionar os itens frequentes e infrequentes.

$$supp(X \rightarrow Y) = \frac{\rho(X \cup Y)}{N} \quad (1)$$

$$conf(X \rightarrow Y) = \frac{\rho(X \cup Y)}{\rho(X)} \quad (2)$$

O valor da Confiança é utilizado para a geração de regras de associação (Etapa 2), compondo regras do tipo $[VetorPixel = 34{,}50{,}40{,}71 \rightarrow class = 1]$, deste modo, o antecedente X é tudo que está à esquerda da implicação ($\rightarrow$), e Y é toda a sentença que está a direita.

2.2 Modelo Baseado em Suporte e Confiança

O modelo Suporte/Confiança, por ser o modelo típico de regra de associação utiliza os valores mínimos de Suporte Equação 1 e Confiança Equação 2. Esses valores, são informados pelo usuário para gerar regras de associação que consiste em um Suporte e Confiança maiores ou iguais aos mínimos informados. Assim o modelo Suporte/Confiança compõe as Etapas 1 e 2 do processo de geração de regras de associação. O modelo limita

o número de itens frequentes (Etapa 1) e o número de regras de associação (Etapa 2), baseado nos valores de Suporte e Confiança informados, funcionando como métricas por serem capazes de eliminar determinadas regras. Experiências apontadas em WITTEN et al. (2016) demonstraram que a mineração de bases de dados reais pode levar à concepção de milhares de regras de associação. O Modelo Suporte/Confiança tem recebido muitas críticas por selecionar e eliminar suas regras com apenas duas métricas, Suporte e Confiança. O quantitativo de regras geradas pelo modelo geralmente é amplo, dificultando a análise por parte de quem a utiliza.

2.3 Medidas de Interesse: Objetivas e Subjetivas

Para resolver algumas das limitações do modelo Suporte/Confiança, medidas de interesse alternativas têm sido propostas com o intuito de identificar as regras de associação que são, de fato, úteis dentre as muitas que podem ser descobertas em mineração de imagens.

Existem medidas de interesse subjetivas que, na prática, consideram principalmente o conhecimento de um profissional para determinar a qualidade da de um padrão associativo. E medidas de interesse objetivas que utilizam índices estatísticos para avaliar a força de uma regra de associação.

Estudo abordado por GENG e HAMILTON (2006), demonstra mais de 20 medidas de interesse que podem ser alternativas para Suporte e Confiança, ou servirem como ferramentas complementares para construção de novas medidas de interesse.

Destaca-se as métricas *Conviction*, Equação 3 proposta por BRIN et al. (1997), *Gini Index*, Equação 4, *Mutual Information*, Equação 5 exibidas em TAN; KUMAR e SRIVASTAVA (2004), *Hyper-Confidence,* Equação 6 em HAHSLER e HORNIK (2007) e Imbalance Ratio Equação 7, Kulczynski Equação 8 demonstradas em WU; CHEN e HAN (2010).

As métricas apresentadas nas Equações 3, 4, 5, 6, 7 e 8, são aplicadas a regras $X \rightarrow Y$, tal que, X é o antecedente e Y o consequente. Com valores entre 0 e 1, com exceção da Equação 3 que pode ter valores entre 0 e ∞.

$$conviction(X \rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \rightarrow Y)} = \frac{P(X)P(\overline{Y})}{P(X \cap \overline{Y})} \quad (3)$$

Gini Index, Equação 4, entropia de medições quadráticas, é uma medida de dispersão

podendo variar o seu valor entre 0 e 1, e assume o valor de 0 quando o antecedente X e o consequente Y são estatisticamente independentes e 0,5 para uma perfeita correlação. Esta medida não é simétrica portanto, quantifica a implicação real de $X \rightarrow Y$.

$$gini(X \rightarrow Y) = P(X)\left[P(Y|X)^2 + P(\overline{Y}|X)^2\right] + P(\overline{X})\left[P(B|\overline{X})^2 + P(\overline{Y}|\overline{X})^2\right] - P(Y)^2 - P(\overline{Y})^2 \quad (4)$$

Mutual Information, Equação 5, mede o ganho de informações para Y fornecido por X. Quantificando a informação compartilhada de uma relação $X \rightarrow Y$, avaliando sua similaridade. Quando 0, X é independente de Y, 1 para dependentes.

$$M(X \rightarrow Y) = \frac{\sum_{i\in\{X,\overline{X}\}}\sum_{j\in\{Y,\overline{Y}\}}\frac{c_{ij}}{n}log\frac{c_{ij}}{c_i c_j}}{min\left(-\sum_{i\in\{X,\overline{X}\}}\frac{c_i}{n}log\frac{c_i}{n}, -\sum_{j\in\{Y,\overline{Y}\}}\frac{c_j}{n}log\frac{c_j}{n}\right)} = \frac{\sum_{i\in\{X,\overline{X}\}}\sum_{j\in\{Y,\overline{Y}\}}P(i\cap j)log\frac{P(i\cap j)}{P(i)P(j)}}{min\left(-\sum_{i\in\{X,\overline{X}\}}P(i)log\,P(i), -\sum_{j\in\{Y,\overline{Y}\}}P(j)log\,P(j)\right)} \quad (5)$$

Hyper-Confidence, Equação 6, usada para filtrar ou classificar regras de associação mineradas. Onde regras falsas são problemáticas. Um nível de confiança de, por exemplo, maior ou igual a 0,95 indica que há apenas 5% de chance de a regra ser gerada aleatoriamente.

$$hyper - conf(X \rightarrow Y) = 1 - P[C_{XY} \geq c_{XY}|c_X, c_Y] \quad (6)$$

Imbalance Ratio, Equação 7, mede o grau de desequilíbrio entre dois eventos em que o antecedente X e o consequente Y estão contidos em uma ocorrência. A proporção é próxima de 0 se as probabilidades condicionais forem semelhantes (ou seja, muito equilibradas) e próxima de 1 se forem muito diferentes. 0 indica uma regra equilibrada, geralmente desinteressante.

$$IB(X \rightarrow Y) = |P(X|Y) - P(Y|X)|\}\{P(X|Y) + P(Y|X) - P(X|Y)P(Y|X)) = |supp(X) - supp(Y)|supp(X) + supp(Y) - supp(X \cup Y) \quad (7)$$

Kulczynski, Equação 8, preferência por padrões inclinados leva em consideração a relação $X \rightarrow Y$ e $Y \rightarrow X$, assim descobrindo a inclinação do antecedente X para o seu

consequente Y. Quando 0,5 significa neutro e geralmente desinteressante.

$$kulc(X \rightarrow Y) = \frac{1}{2}\big(conf(X \rightarrow Y) + conf(Y \rightarrow X)\big) = \frac{1}{2}\left(\frac{supp(X \cup Y)}{supp(X)} + \frac{supp(X \cup Y)}{supp(Y)}\right) \quad (8)$$

$$= \frac{1}{2}\big(P(X|Y) + P(Y|X)\big)$$

3. MÉTODOS

Nosso método é constituído de um Framework com cinco fases (Figura 2). Composto por técnicas de tratamento digital de imagens alinhado a uma estratégia de manipulação de dados, nossa abordagem visa aplicar técnicas de mineração de dados valorizando conhecimento do especialista para segmentar imagens aéreas de plantações (citros).

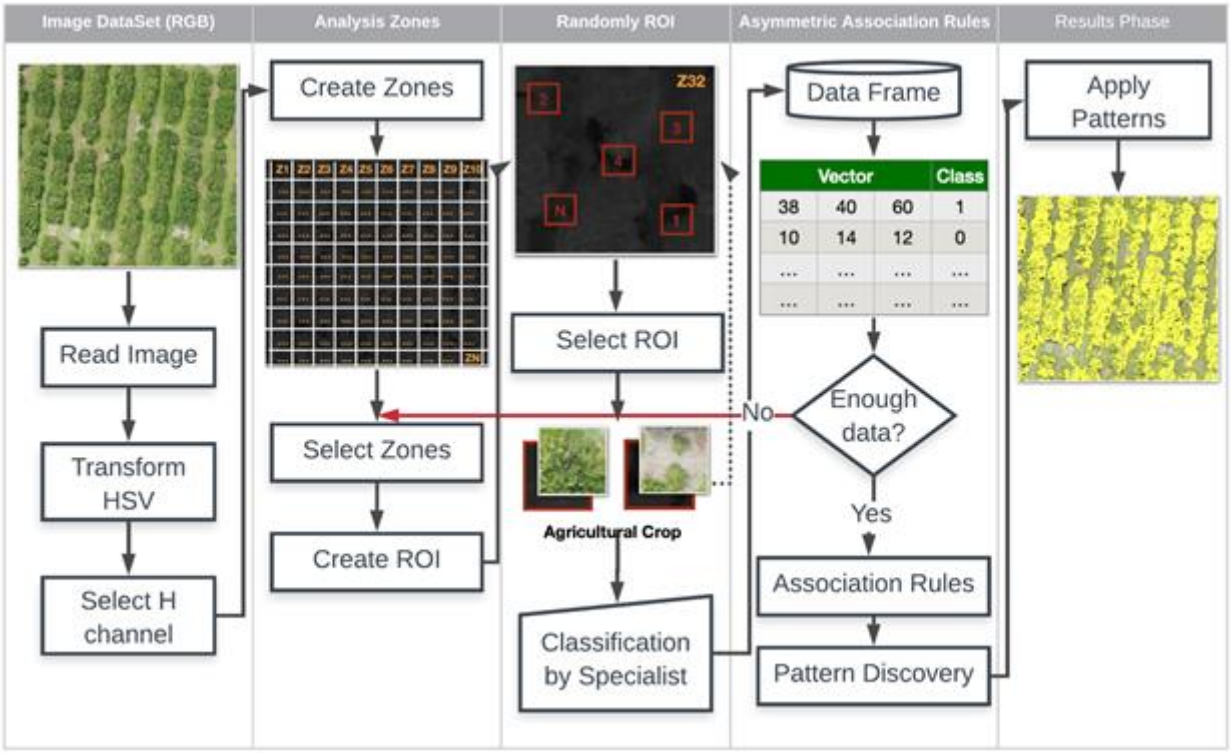

Figura 2. Fluxograma do Framework de 5 Fases.

3.1 Conjunto de Dados de Imagens

Nesta fase a imagem é carregada e transformada para o sistema de cores HSV formadas pelas componentes Hue (matiz), Saturação e Valor. Selecionando o canal H deste sistema. O principal objetivo é diminuir ruídos e eliminar a interferência de sombras nas imagens das plantações.

3.2 Zonas de Análise

A criação de zonas em nossa abordagem é fundamental para garantir que será adquirido o mínimo de amostras de cada parte da imagem, assim podemos eliminar possíveis vieses no momento de rotulação das classes de interesse na imagem.

Cada uma das zonas é definida dentro de um gradeado, onde divide-se a imagem com dimensão D (altura x largura da imagem) por um valor k. Por exemplo, se tivermos uma imagem de dimensão D = 1000 x 250 pixels e k = 10, tem-se um gradeado onde cada zona tem 100 x 25 pixels, dando um número de 100 zonas.

A seleção das zonas que passarão para a próxima etapa, se dá, nas zonas que tiverem mais de 70% dos pixels da mesma cor (caracterizando a ocorrência de plantação). Um detalhe, que em cada uma das zonas de análise, cortes aleatórios serão criados na fase seguinte para rotulação pelo especialista.

Ressalta-se que os valores de k = 10 e o limite de 70% ou mais dos pixels da mesma cor para selecionar a zona de análise da imagem, foram determinados experimentalmente.

É possível que diferentes valores podem ser utilizados para personalizar ajustes em alguns conjuntos de dados de imagens de plantações. Um detalhe a ser observado é que quanto menor o número de zonas, maior deverá ser o número de recortes da fase de Recorte Aleatório.

3.3 Recorte Aleatório

Para realizar a seleção criamos N recortes em cada uma das zonas. O valor padrão de N é definido por meio das características da imagem carregada na primeira etapa (Figura 2). Por meio de experimentos, definimos um valor de $N = 10$ como padrão para imagens com até duas classes (Figura 3).

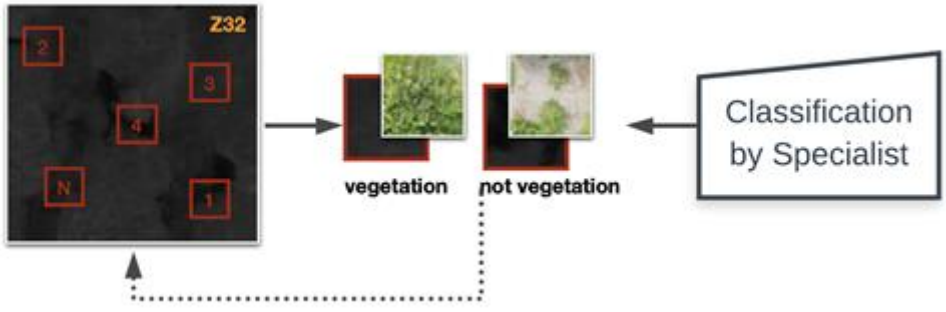

Figura 3. Fluxograma da etapa de Recorte Aleatório.

Para cada uma das zonas Z, é criado N recortes randomicamente. Esses recortes são apresentados para o especialista para que ele classifique (*rotule*) o recorte em plantação ou solo. Isso ocorre recursivamente até que N chegue ao número máximo estabelecido para cada uma das zonas. No momento da classificação do especialista, ele pode descartar recortes, caso ele não consiga definir qual é a classe mais adequada. Quando ocorrer descarte, um outro recorte é colocado no lugar até que o número estabelecido em N seja alcançado.

Os dados resultantes das classificações do especialista são transformados em um vetor de características e armazenado em um Data Frame (Dados em Memória) juntamente com a sua definição de classe.

3.4 Regras de Associação Assimétricas

Nesta fase é aplicado o algoritmo de *Image Mining* utilizando os dados da fase anterior. Um Data Frame é construído de maneira incremental com a classificação aplicada por meio do especialista. Por ser incremental o Data Frame fica mais rico de dados para cada imagem rotulada por meio da abordagem, conforme apresentado na Figura 4.

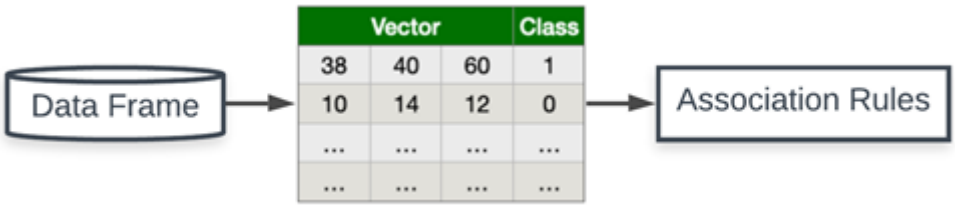

Figura 4. Fluxograma de Criação da Regras de Associação.

Para cada um dos N recortes das zonas Z, os dados são transformados em um vetor que é armazenado em um Data Frame, seguido da rotulação definida por meio do especialista. A cada classificação dos rótulos o Data frame recebe esses dados que são armazenados de maneira acumulativa. Até que o *Data Frame* tenha dados o suficientes, novas amostras podem ser obtidas enquanto existir zonas Z disponíveis.

Nossa abordagem utiliza um algoritmo de regras de associação (RA) customizado. Diferente dos algoritmos tradicionais de padrões associativos, no *Image Mining* adotamos uma solução que utiliza seis métricas assimétricas probabilísticas para identificação dos padrões (Seção 2.3). A utilização das métricas em nossas abordagens são aplicadas nos passos do nosso algoritmo de regras assimétricas conforme descrito na Figura 5.

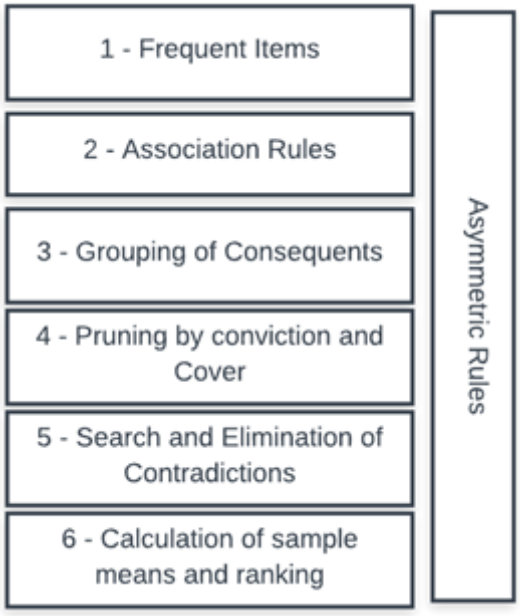

Figura 5. Algoritmo *Image Mining* para criação de Regras de Associação Assimétricas.

Na etapa 1 da Figura 5, após o carregamento dos dados por meio do Data Frame, o valor do *Suporte* e *Confiança* é calcula por $\frac{1}{N}$ onde N, onde N é o número de elementos no banco de dados. Assim será atribuído o menor valor possível de Suporte para o conjunto de dados, desta forma, é obtido o máximo de itens frequentes. Diferente dos algoritmos tradicionais de RA onde o conjunto de itens frequentes e regras são definidas por um mínimo Suporte e Confiança, nossa abordagem define um limite superior utilizando a métrica Convicção. O limite superior é informado por meio do usuário do algoritmo. Baseado na métrica Convicção que tem sua variação entre 0 e ∞.

A partir do conjunto de itens frequentes gerados na etapa 1, conjuntamente com o valor da *Confiança*, é gerado o conjunto de regras de associação na etapa 2 da Figura 5. No algoritmo original a etapa é calculada apenas a *Confiança* para cada uma das regras, no entanto, é adicionado em nossa abordagem, o cálculo para todos os padrões associativos de métricas adicionais (*Hyper-Confidence, Mutual Information, Imbalance Ratio, Kalczynski* e *Gini Index*).

Na etapa 3 da Figura 5 é realizado o *Grouping of Consequent*, que utiliza o limite superior e limite inferior determinados. Aqui, todos os padrões de associação que não respeitem os valores estabelecidos são eliminados. As regras restantes do processo de poda são agrupadas. Todas as regras $X \rightarrow Y$ são separadas em subconjuntos de padrões. O limite inferior da Convicção é determinado pela média aritmética da confiança de cada um dos subconjuntos de consequentes.

Cada subconjunto de padrões é organizado por seu valor de consequente Y. Dividindo o conjunto de regras resultante da etapa anterior (2) em N subconjuntos. A regra do topo de cada subconjunto, é selecionada e verificada com as outras regras de maneira recursiva, em busca de regras cobertas por outras, até que todas as regras e subconjuntos sejam verificados.

Baseado no estudo de TOIVONEN et al. (1995), que propôs a ideia de eliminar padrões redundantes usando cobertura de regras estruturais, tem-se a etapa 4 da Figura 5 que é *Pruning by conviction and Coverage*. Essa etapa (4) seleciona em cada subconjunto todas as regras com os primeiros antecedentes X idênticos. Como exemplo, o subconjunto [*Class=1*], a regra [*38,40,60 → Class=1*], exibe três antecedentes a segunda com um único antecedente [*38,38,38 → Class=1*]. As regras com seus primeiros antecedentes idênticos, são selecionadas e comparadas por apresentarem informações similares, independente do segundo antecedente.

Isso é feito para eliminar as regras com menor média entre *Hyper-Confidence* e *Mutual Information*. A primeira métrica garante que sejam escolhida as regras com menor chance de serem geradas de maneira aleatória. Já a segunda, mede o ganho de informação do consequente Y fornecido por meio do antecedente X. Assim, é eliminada a regra que já possui o valor de pixel de maneira repetitiva (por exemplo, 38), por já estar presente em outra regra associativa e com a mesma classe final (*Class = 1*).

Já a etapa 5 da Figura 5, denominada *Search and Eliminate of Contradictions*. Nessa etapa a contradição de sentido é determinada, por exemplo, nas regras 1 e 2 (Tabela 1). Em ambas as situações não é trivial definir qual padrão poderia ser eliminado. Utilizando a média entre as métricas de Confiança e a métrica de Kalczynski, que verifica os padrões de inclinação levando em consideração a relação de $X \rightarrow Y$ e $Y \rightarrow X$. Aplicando essas métricas para escolher as regras em contradições de sentido, seleciona-se a regra com maior valor de inclinação e eliminando a outra.

Tabela 1. Exemplo de Contradições.

id	Antecedente X	Consequente Y
01	38,39,40 →	Class=1
02	Class=0 →	38,39,40

Finalmente, tem-se a etapa 6 da Figura 5, denominada *Calculation of sample means and ranking.* A Figura 6, apresenta um gráfico de coordenadas paralelas com as regras

geradas por meio da abordagem da etapa 6. As linhas em cinza apresentam as regras do conjunto de dados de teste e em laranja a regra ideal.

A regra ideal é composta de valores das métricas objetivas (*Hyper-Confidence, Gini Index, Mutual Information, Imbalance Ratio* e *Kalczynski*) com os respectivos valores [0.95, 0.3, 1, 1, 0.6] que em conjunto definem a regra com maior potencial para o contexto da aplicação desejada. As linhas em cinza, apresentam o comportamento das regras em relação as métricas utilizadas na abordagem.

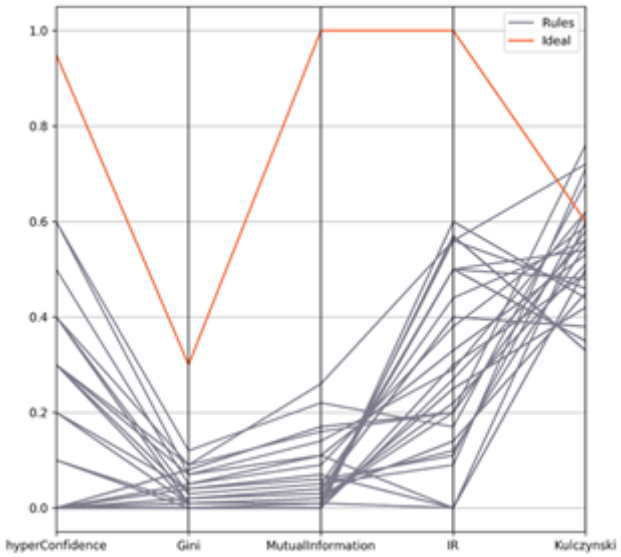

Figura 6 Exemplo do comportamento das regras.

O cálculo da distância das regras e ranqueamento de padrões identificados, é calculada pela Diferença das Médias Amostrais (*DMA*) introduzido por ROTHSCHILD e STIGLITZ (1970). Calculada entre *ideal* e as regras descobertas no conjunto de dados. Equação 9 apresenta a fórmula utilizado nesse estudo para o cálculo das distâncias das regras.

$$DMA = \bar{X}_i - \bar{M} \quad (9)$$

O $\bar{X}_i$ indica a média das métricas *Hyper-Confidence, Gini Index, Mutual Information, Imbalance Ratio* e *Kulczynski*. A média aritmética da regra ideal é definida por $\bar{M}$. Quanto mais próximo de 0 for o *DMA* melhor será o ranqueamento da regra na nossa abordagem.

Os padrões identificados com melhor *DMA* são separados e filtrados, referentes às classes de interesse (vegetação, solo). Com essa abordagem de regras de associação assimétricas é possível reunir os padrões mais relevantes para cada uma das classes e

concatená-las, definindo assim, um conjunto reduzido de regras de associação que definem as classes baseada no conhecimento do especialista.

Os padrões resultantes do algoritmo de associação são aplicados à imagem original, segmentando a plantação. Embora os dados resultantes estejam com os valores dos pixels do canal H, utilizamos a imagem original para plotar, respeitando as mesmas coordenadas entre imagem original e canal H do HSV. Facilitando o processo de visualização conforme apresentado na Figura 7.

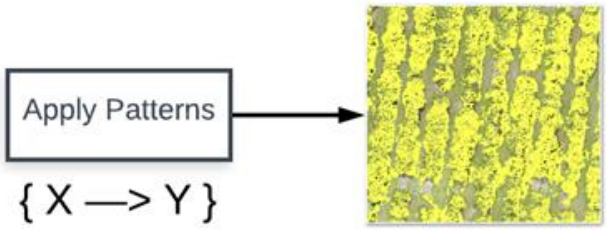

Figura 7 Regras de Associação Assimétricas.

4. RESULTADOS E DISCUSSÕES

A abordagem proposta foi utilizada em seis conjuntos de dados com características distintas. No entanto, aplicamos o que foi desenvolvido na abordagem e no algoritmo de padrões associativos na plantação de citros, conforme dados apresentados na Tabela 2.

4.1 Conjunto de Imagens

A Figura 8 é uma amostra dos mosaicos de imagens de plantações utilizados e cujas características de dimensões, ângulos e coloração estão descritos na Tabela 2.

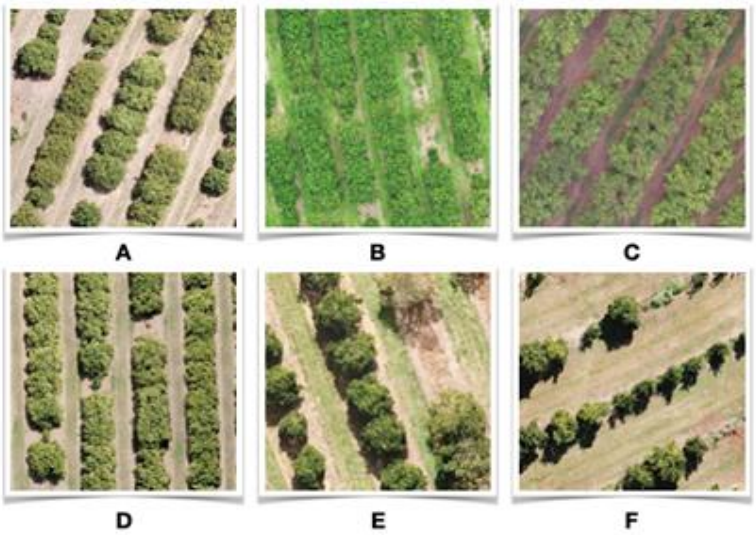

Figura 8. Amostra dos Mosaicos de Plantações.

Tabela 2. Conjunto de Imagens de Plantações.

id	Plantação	Dimensões
A	Citros	4634 x 3106
B	Citros	13865 x 13928
C	Citros	8648 x 3819
D	Citros	2878 x 6101
E	Citros	3626 x 2957
F	Citros	4859 x 6642

4.2 Experimentos

Com N=10 como o números de amostras classificadas pelo especialista, a abordagem conseguiu capturar o número ideal de pixels que representa as classes de terreno e plantação. O processo de utilização do algoritmo de associação assimétrico filtrou os valores mais adequados para o grupo de pixel alvo.

Conforme é possível ver na Figura 9, os mosaicos A, D, F e C tiveram uma segmentação eficiente da plantação, conforme destacado em amarelo, demostrando que com N=10 a abordagem proposta consegue realizar o processo de segmentação. Diferente da abordagem baseada em modelos (Aprendizagem de máquina), a nossa é capaz de, com um número reduzido de amostras, realizar o processo de segmentação.

Destacamos ainda que a abordagem proposta associa os pixels de interesse com a classe pretendida, podendo obter bons resultados em variados contextos, por se basear no conhecimento de especialistas.

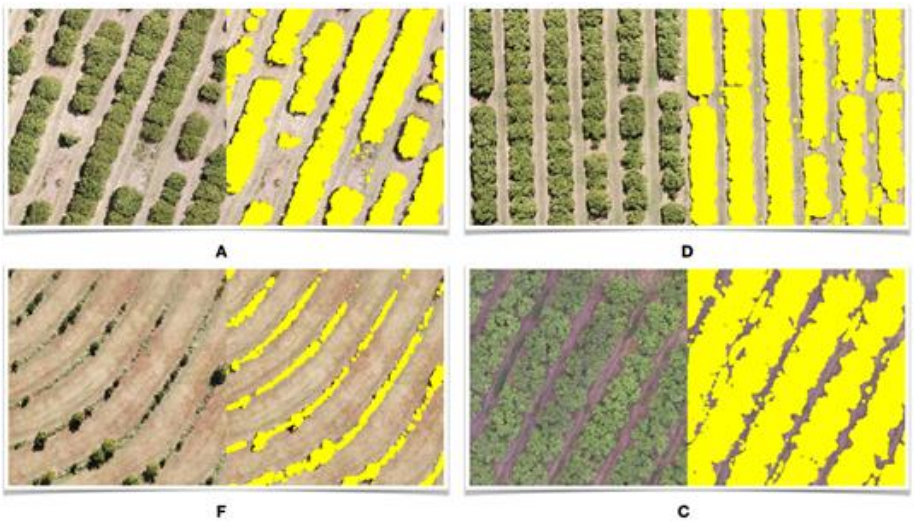

Figura 9. Melhores Resultados.

Outros resultados apresentados apontam para cuidados que devem ser tomados ao utilizá-la. No mosaico B da Figura 10 observamos que quando o N < 10 o processo de segmentação pode ficar prejudicado. Isso, devido à falta de dados que interferem na escolha dos grupos de pixels que melhor definem a classe alvo.

Observamos também que quando o mosaico não é homogêneo como o mosaico E da Figura 10, que apresenta plantas com uma grande diferença de tamanho, a abordagem tem dificuldades em definir o padrão ideal para mosaicos não homogêneos, assim segmentando de maneira confusa conforme o mosaico E e não tão satisfatório no mosaico E-2, ambos da Figura 10.

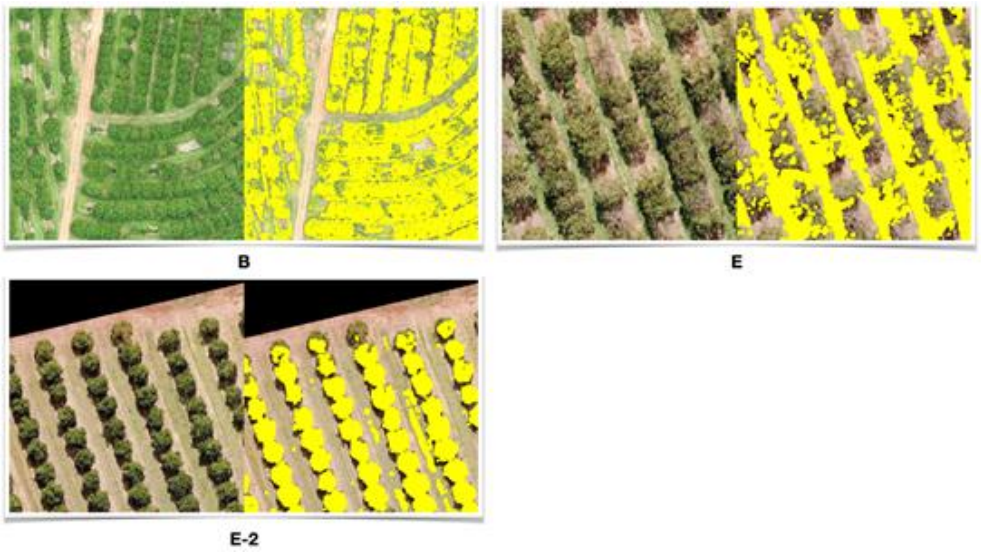

Figura 10. Outros Resultados.

4.3 Padrões de Interesses Encontrados

Os padrões de interesse encontrados através dos filtros e usados em nossa abordagem, relacionam diretamente os valores de pixels da imagem em avaliação com a classe observada.

Tabela 3. Amostra de Padrões Associativos Assimétricos.

id	Antecedente X	Consequente Y
01	38,39,40 →	Class=1
02	Class=0 →	38,39,40

A Tabela 3 apresenta exemplos de padrões encontrados para classe 1 (plantação) e para classe 0 (solo). Desta forma os valores de pixel predominantes são associados a classe de referência.

5. CONSIDERAÇÕES FINAIS

Os experimentos realizados tiveram como ponto central sete mosaicos de plantações de citros, onde tivemos sucesso na identificação das classes desejadas. Nossa solução é composta por um Framework e um aprimoramento do algoritmo de associação, que melhoram a seleção de padrões associativos quando aplicados à mineração de imagens.

O principal diferencial do nosso estudo é que o conhecimento do especialista é valorizado, sendo a principal forma de comparações para que o algoritmo possa identificar os padrões. Outro ponto forte é que não há necessidade de ajustes de rotina no modelo, como no caso das árvores de decisão.

Nós aplicamos a técnica proposta a dados reais e conseguimos validar os pontos fortes e fracos de nossa solução. Ressaltamos que nossa solução se limita a situações com duas classes de interesse, nesse caso específico, seriam plantação e solo. Percebemos que nosso estudo pode ser aplicado a estudos relacionados a plantios de diferentes formatos, identificando falhas na plantação, podendo ser utilizado para planejamento urbano e outras soluções que utilizam imagens aéreas como ponto de partida.

Acreditamos que uma evolução viável para nosso estudo é o desenvolvimento de produtos inovadores que utilizem nosso embasamento teórico e experimental para criar soluções para a agricultura de precisão.

REFERÊNCIAS

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Mining association rules between sets of items in large databases. **Acm sigmod record**. [S.l.], v. 22, n. 2, p. 207–216, 1993.

AGRAWAL, R., SRIKANT, R., et al. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, **VLDB. [S.l.: s.n.],** 1994. v. 1215, p. 487–499, 1994.

ASRIN, F., SAIDE, S., RATNA, S., WENDA, **A. Knowledge data discovery (frequent pattern growth): The association rules for evergreen activities on computer monitoring**. SPRINGER. International Conference on Intelligent and Fuzzy Systems. [S.l.], p. 807–816, 2020.

BRIN, S., MOTWANI, R., ULLMAN, J. D., TSUR, S. Dynamic itemset counting and implication rules for market basket data. SIGMOD Rec., ACM, New York, NY, USA, v. 26, n. 2, p. 255–264, 1997.

BUXTON, E. K., VOHRA, S., GUO, Y., FOGLEMAN, A., PATEL, R. Pediatric population health analysis of southern and central Illinois region: A cross sectional retrospective study using association rule mining and multiple logistic regression. **Computer Methods and Programs in Biomedicine**, v. 178, 145 – 153, 2019.

DESHMUKH, J., BHOSLE, U. **Image mining using association rule for medical image dataset. Procedia Computer Science** v. 85, 117 – 124. International Conference on Computational Modelling and Security, 2016.

DEY, NILANJAN A, W. B. A., CHAKRABORTY, S., BANERJEE, S., SALEM, M. A., AZAR, A. T. Image mining framework and techniques a review. **International Journal of Image Mining** 1, 45–64, 2015.

FATMA, S. N., NASHIPUDIMATH, M. **Image mining using association rule**. World Congress on Information and Communication Technologies. IEEE, 587–593, 2011.

GENG, L., HAMILTON, H. J. **Interestingness measures for data mining: A survey.** ACM Comput. Surv. v. 38, 2006.

GUREVICH, I. B., YASHINA, V. V. **Descriptive image analysis: Genesis and current trends. Pattern Recognition and Image Analysis**, v. 27, 653–674, 2017.

HAHSLER, M., HORNIK, K. New probabilistic interest measures for association rules. Intell. Data Anal. 11, 437–455, 2007.

HAN, J., KAMBER, M., PEI, J. The Morgan Kaufmann Series in Data Management Systems. Third Edition. Morgan Kaufmann, Boston, 279 – 325, 2012.

HAN, J., PEI, J., YIN, Y., MAO, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, v. 8, 53–87, 2004.

HSU, W., LEE, M. L., ZHANG, J. Image mining: Trends and developments. **Journal of intelligent information systems**. v. 19, 1, 7–23, 2002.

KAPLUN, D. I., VOZNESENSKIY, A. S., KLIONSKIY, D. M., GEPPENER, V. V., SERZHENKO, F. L. **Study of spatiotemporal processing algorithms for video and image processing. Pattern Recognition and Image Analysis**. v. 27, 686–694, 2017.

KHAN, S., SOOMRO, T. R., ALAM, M. M. Application of Image Processing in Detection of Bone Diseases Using X-rays. Pattern Recognition and Image Analysis. v. 30, 97–107, 2020.

KUMAR, N., KAUR, N., GUPTA, D. Red Green Blue Depth Image Classification Using Pre-Trained Deep Convolutional Neural Network. Pattern Recognition and Image Analysis. v.30, 382–390, 2020.

LAKSHMI, K., VADIVU, G. **Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Computer Science**. 7th International Conference on Advances in Computing & Communications, ICACC, 22-24, 2017.

MUNAWAR, H. S., ZHANG, J., LI, H., MO, D., CHANG, L. **Mining multispectral aerial images for automatic detection of strategic bridge locations for disaster relief missions**. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 189–200, 2019.

OSADCHIY, T., POLIAKOV, I., OLIVIER, P., ROWLAND, M., FOSTER, E. Recommender system based on pairwise association rules. **Expert Systems with Applications**. v. 115, 535 – 542, 2019.

ROTHSCHILD, M., STIGLITZ, J. E. Increasing risk: I. a definition. **Journal of Economic Theory**. v. 2, 225 – 243, 1970.

TAN, P.-N., KUMAR, V., SRIVASTAVA, J. Selecting the right objective measure for association analysis. **Information Systems**, v. 29, n. 4, p. 293 – 313, 2004.

TAN, P.-N., STEINBACH, M., KUMAR, V. **Introduction to data mining: Pearson Addison** Wesley. Boston, 2005.

TOIVONEN, H., KLEMETTINEN, M., RONKAINEN, P., HÄTÖNEN, K., MANNILA, H. Pruning and grouping discovered association rules, 1995.

WITTEN, I. H., FRANK, E., HALL, M. A., PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.

WU, T., CHEN, Y., HAN, J. Re-examination of interestingness measures in pattern mining: a unified framework. Data Mining and Knowledge Discovery. v. 21, 371–397, 2010.

XU, C., BAO, J., WANG, C., LIU, P. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in china. **Journal of Safety Research.** v. 67, 65 – 75, 2018.

ZAHRADNIKOVA, B., DUCHOVICOVA, S., SCHREIBER, P. Image mining: review and new challenges. **International Journal of Advanced Computer Science and Applications**. v. 6, 242–246, 2015.

ZHANG, Y., NIU, L. A Substructure Segmentation Method of Left Heart Regions from Cardiac CT Images Using Local Mesh Descriptors, Context and Spatial Location Information. **Pattern Recognition and Image Analysis.** v. 29, 230–239, 2019.