LUIZ OTÁVIO LAMARDO ALVES SILVA

CLASSIFICAÇÃO VISUAL DE MUDAS DE PLANTAS
ORNAMENTAIS: ANÁLISE DA EFICÁCIA DE TÉCNICAS DE SELEÇÃO
DE ATRIBUTOS

São Paulo

2014

LUIZ OTÁVIO LAMARDO ALVES SILVA

CLASSIFICAÇÃO VISUAL DE MUDAS DE PLANTAS
ORNAMENTAIS: ANÁLISE DA EFICÁCIA DE TÉCNICAS DE SELEÇÃO
DE ATRIBUTOS

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Ciências.

São Paulo

2014

LUIZ OTÁVIO LAMARDO ALVES SILVA

CLASSIFICAÇÃO VISUAL DE MUDAS DE PLANTAS
ORNAMENTAIS: ANÁLISE DA EFICÁCIA DE TÉCNICAS DE SELEÇÃO
DE ATRIBUTOS

Dissertação apresentada à Escola
Politécnica da Universidade de São
Paulo para obtenção do título de
Mestre em Ciências.

Área de Concentração:
Engenharia de Computação

Orientador: Prof. Livre-Docente Carlos
Eduardo Cugnasca

São Paulo

2014

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 28 de janeiro de 2014.

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Silva, Luiz Otavio Lamardo Alves

Classificação visual de mudas de plantas ornamentais: análise da eficácia de técnicas de seleção de atributos / L.O.L.A.

Silva. – versão corr. -- São Paulo, 2014.

64 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Produutos agrícolas 2.Impressão visual I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Dedico este trabalho
aos meus pais pelo apoio e
incentivo que me proporcionaram.

AGRADECIMENTOS

Ao meu orientador, Prof. Livre Docente Carlos Eduardo Cugnasca, pela orientação e dedicação na condução deste trabalho e, sobretudo, pela confiança e liberdade que me propiciou durante a realização do trabalho, sem as quais provavelmente não haveria chegado até aqui.

À minha família que é meu porto seguro e que me deu incentivo, apoio e suporte para sempre seguir em frente. Em especial, à minha mãe Cecília e meu pai Mario, que contribuíram fortemente na minha formação pessoal e profissional. Pelas conversas madrugadas dentro e pelos conselhos à beira da mesa.

Ao mestrando Marcelo Li Koga juntamente com meu caro amigo Laio Burim Vilas Bôas pela contribuição, incentivo e apoio ao longo desses dois anos.

A Prof^a. Anna Helena Reali Costa, pela dedicação, atenção e disposição em me receber sempre que precisei de seus conselhos e conhecimento.

À minha querida Lara Tasso Guaraldo, pelo carinho, amor e compreensão tão bem-vindos durante todo o trabalho.

Aos companheiros de pesquisa do LAA, em especial ao Marcelo Barros, Allan, Wilian, Michel, Ivairton e outros que contribuíram diretamente e indiretamente para a realização deste trabalho. Assim como a equipe de colaboradores do laboratório e do PCS.

Aos amigos, por contribuírem com os bons momentos de descanso e com o enriquecimento da minha formação nas atividades extracurriculares.

Esse trabalho foi realizado com apoio do CNPq.

In God we trust, all others bring data.
(William Edwards Deming)

RESUMO

A automação da classificação visual de produtos vem ganhando cada vez mais importância nos processos produtivos agrícolas. Isto posto, uma das principais dificuldades encontradas por produtores de flores e plantas ornamentais é garantir o crescimento homogêneo de suas plantas. Nesse cenário, as mudas utilizadas para gerar as plantas são importantes uma vez que se pode estimar seu potencial de crescimento através de uma inspeção visual. Sendo assim, um sistema de visão computacional pode ser empregado para automatizar essa tarefa. Porém, diferentemente de indústrias tradicionais, a indústria agrícola apresenta grande variabilidade entre os produtos analisados. Técnicas de aprendizado de máquina supervisionado conseguem avaliar um conjunto de atributos referentes ao objeto inspecionado para classificá-lo corretamente, de forma a lidar tanto com a variabilidade dos produtos em inspeção quanto com a incorporação do conhecimento de especialistas pelo sistema. A definição do conjunto de atributos a ser extraído das imagens dos produtos é de extrema importância, pois é ele quem fornece toda a informação utilizada no sistema. Um conjunto com diversos atributos assegura que toda a informação necessária é capturada, porém atributos irrelevantes ou redundantes podem prejudicar o desempenho dos classificadores. Técnicas de seleção de atributos podem ser utilizadas para equilibrar essas necessidades. O objetivo do trabalho foi o de avaliar a eficácia dessas técnicas para a classificação de mudas de violeta. Vinte e seis parâmetros foram extraídos de seiscentas imagens rotuladas em quatro níveis de qualidade. Em seguida, os desempenhos de seis classificadores foram comparados considerando-se um universo de subconjuntos gerados por quatro técnicas de seleção de atributos. Os resultados mostraram que essas técnicas são realmente vantajosas, gerando ganhos de até 8,8% nas taxas de acertos e ao mesmo tempo reduzindo de 26 para 11 o número médio de atributos utilizados. O classificador *Logistic Regression* associado ao subconjunto gerado pelo Chi-quadrado foi o que apresentou melhor desempenho global, atingindo 80% de acerto. O classificador *Random Forest* ficou em segundo lugar, porém se mostrou menos sensível a seleção de atributos.

Palavras-chave: Visão computacional. Aprendizado de máquina. Seleção de atributos. Mudas. Flores. Produtos Agrícolas. Seleção. Separação.

ABSTRACT

The automation of visual classification of products is gaining more importance in agricultural production processes. That said, one of the main difficulties encountered by ornamental plants and flowers producers is to ensure homogeneous growth of their plants. In this scenario, the seedlings used to grow the plants are very important since it is possible to estimate their growth potential by means of a visual inspection. Therefore, a computer vision system can be used to automate this task. Unlike traditional industries, the agricultural industry shows great variability among the products inspected. Supervised machine learning techniques can evaluate an attribute set representing the inspected object in order to correctly classify it, making it possible not only to deal with the variability of the inspected products but also with the incorporation of experts knowledge into the system. The definition of the attribute set to be extracted from the images of the products is of utmost importance, as it is it that provides all information used by the system. A set with several attributes ensures that all necessary information is captured; however irrelevant or redundant attributes can affect the performance of classifiers. Attribute selection techniques can be used to balance these needs. The aim of this study was then to evaluate the effectiveness of these techniques regarding the classification of African violet seedlings. Twenty-six parameters were extracted from six hundred images, labeled into four quality groups. Then, the performances of six classifiers were compared by considering the universe of subsets generated by four attribute selection techniques. The results showed that these techniques are indeed advantageous, generating gains of up to 8.8% in accuracy rate while reducing from 26 to 11 the average number of attributes used. Logistic Regression classifier, associated with the subset generated by the Chi-squared filter showed the best overall performance, achieving 80 % accuracy. Random Forest was second, but was less sensitive to attribute selection

Keywords: Computer vision. Machine learning. Attribute selection. Cuttings. Flowers. Agricultural products. Sorting. Selection.

LISTA DE ILUSTRAÇÕES

Figura 1 - Dois exemplos de aplicação tradicional de sistemas de visão computacional na indústria. Superior: bebidas engarrafadas. Inferior: peças metálicas.	20
Figura 2 - Diagrama das etapas do funcionamento de um sistema de visão computacional industrial. A tomada de decisão é feita a partir de regras explícitas fornecidas pelo usuário.	22
Figura 3 - Exemplos de classificações de produtos agrícolas. Na parte superior: mudas de violeta por estaquia. Inferior: batatas lavadas.....	23
Figura 4 – Diagrama das etapas do funcionamento de um sistema de visão computacional que utiliza aprendizado de máquina supervisionado ao realizar a inspeção. O sistema precisa ser previamente treinado com exemplos para inferir as regras de classificação.	25
Figura 5 – Representação gráfica do processo de decisão do kNN.	26
Figura 6 – Função logística.....	27
Figura 7 – Esquema de grafo orientado com camadas totalmente conectadas.	29
Figura 8 – Exemplo de árvore de decisão sobre sucesso na carreira.	29
Figura 9 – Exemplos de imagens das mudas classificadas nos diversos níveis em ordem decrescente de qualidade, sendo (a) um exemplo de muda da melhor qualidade e (d) da pior.	41
Figura 10 – Esquema do ambiente de captura utilizado para obter as imagens das mudas.....	42
Figura 11 - (a) Imagem original. (b) Resultado intermediário, após a binarização. (c) Resultado final, após aplicação de filtros mediana e abertura morfológica.....	43
Figura 12 - Exemplo da obtenção do eixo de simetria longitudinal da muda. O ponto A é o na extrema direita da imagem que pertence a muda, B é o centroide da região destacada e C representa o centroide da muda. A união de B com C estabelece o eixo desejado.....	45
Figura 13 – Visualização dos atributos extraídos das imagens. Os retângulos numerados representam sub-regiões que avaliam a distribuição de massa das mudas.	46

Figura 14 – Gráfico da variação da taxa de acerto obtida ao se variar a quantidade de atributos inclusos no subconjunto. A ordem adotada segue o critério de relevância estabelecido pelo Chi-quadrado.....50

Figura 15 - Gráfico da variação da taxa de acerto obtida ao se variar a quantidade de atributos inclusos no subconjunto. A ordem adotada neste caso segue o critério de relevância estabelecido pelo *Information Gain*.51

Figura 16 - Gráfico da variação da taxa de acerto obtida ao se variar a quantidade de atributos inclusos no subconjunto. Este último gráfico da sequência segue a ordem estabelecida pelo *Gain Ratio*.....52

Figura 17 – Taxas finais de acerto obtidas antes e após a seleção de atributos para todos os casos analisados. Os números acima das barras indicam a quantidade de atributos no subconjunto utilizado.54

LISTA DE ABREVIATURAS E SIGLAS

CFS – *Correlation Based Feature Selector*

COMPAG – *Computers and Electronics in Agriculture*

kNN – *k Nearest Neighbors*

LR – *Logistic Regression*

MLP – *Multilayer Perceptron*

PCA – *Principal Components Analysis*

RF – *Random Forest*

RN – *Redes Neurais*

SUMÁRIO

1	Introdução	13
1.1	Motivação.....	13
1.2	Objetivo.....	15
1.3	Metodologia	15
1.4	Trabalhos correlatos	17
1.5	Organização do texto	19
2	Sistemas de Inspeção Visual e Controle de Qualidade	20
2.1	Sistemas de Visão Computacional Industriais	20
2.2	Particularidades na Inspeção de Produtos Agrícolas.....	21
2.3	Aprendizado de Máquina	23
2.4	Algoritmos de classificação.....	25
2.4.1	k-Nearest Neighbors (kNN).....	26
2.4.2	Naïve Bayes.....	26
2.4.3	Logistic Regression.....	26
2.4.4	Multilayer Perceptron	28
2.4.5	C4.5	29
2.4.6	Random Forest	30
2.5	Considerações finais do capítulo	31
3	Seleção de Atributos	32
3.1	Abordagens para seleção de atributos	33
3.1.1	Filtros	33
3.1.2	Wrappers.....	34
3.1.3	Embebbed.....	35
3.2	Abordagem escolhida no trabalho	35
3.3	Técnicas de seleção de atributos.....	36
3.3.1	Correlation-based Feature Selector (CFS).....	36
3.3.2	Information Gain.....	37
3.3.3	Gain Ratio	38
3.3.4	Chi-Quadrado.....	38
3.4	Considerações finais do capítulo	38

4	Procedimentos Experimentais	40
4.1	Coleta de dados	40
4.2	Binarização	42
4.3	Extração de atributos	43
4.4	Definindo os subconjuntos	46
4.5	Validação cruzada	47
4.6	Considerações finais do capítulo	47
5	Resultados e Discussões	48
5.1	Da filtragem.....	48
5.2	Da avaliação dos subconjuntos	50
5.3	Da comparação entre os classificadores	53
5.4	Considerações finais do capítulo	55
6	Conclusões	56
6.1	Limitações.....	58
6.2	Dificuldades encontradas.....	57
6.3	Trabalhos futuros	59
6.4	Contribuições.....	58
6.5	Considerações finais.....	60

1 INTRODUÇÃO

1.1 Motivação

A produção em grande escala de flores e plantas ornamentais se dá em diversas fases. Produtores estabelecem para cada fase um setor dentro das estufas e, em cada um deles, a disposição e o tratamento das plantas são feitos de maneira homogênea, melhorando a qualidade destas e evitando o surgimento de doenças e pragas. O controle sobre o desenvolvimento das plantas é importante para permitir não só um melhor aproveitamento da área da estufa, como também das demandas comerciais (FAUST; HEINS, 2004).

Diversos fatores, como temperatura média diurna e a relação entre a temperatura diurna e noturna, influenciam a velocidade do desenvolvimento das plantas. Porém em uma estufa, esses fatores são praticamente uniformes para todas as plantas. As mudas são importantes porque representam um fator que, diferentemente dos supramencionados, não apresentam uniformidade no processo produtivo, apresentando grande variabilidade. Visto que garantir a qualidade e uniformidade das plantas é de primordial importância (PARSONS; EDMONDSON; GILMOUR, 2006), o que em geral se faz nos processos produtivos é dividir as mudas em grupos de desenvolvimento semelhantes, por meio da inspeção visual de cada uma delas, que é uma maneira comprovadamente eficaz de se prever o crescimento futuro das plantas (MATTSSON, 1997).

Visão computacional tem sido amplamente utilizada para automatizar tarefas de inspeção visual em muitos setores. Por meio dela é possível obter resultados consistentes na classificação, que deixa de depender do desempenho e atenção dos colaboradores quando realizada manualmente, gerando ganhos de qualidade e produtividade na tarefa da inspeção. Porém, diferentemente de sistemas de visão computacional utilizados em indústrias tradicionais, como a de peças mecânicas ou de bebidas engarrafadas, a classificação visual de produtos agrícolas tem de ser capaz de lidar com a variabilidade inerente de cada exemplar a ser analisado. Além disso, as regras para realizar a categorização muitas vezes são de domínio dos produtores e especialistas da área, e se caracterizam como conhecimento tácito, *i.e.*, difícil de ser transmitido formalmente, o que impõe um desafio para a

automação dessa tarefa. A associação de técnicas de reconhecimento de padrões através de aprendizado de máquina pode auxiliar na superação desses desafios, aumentando a flexibilidade frente a variações dos produtos analisados.

A construção de sistemas computacionais baseados em conhecimento tácito enfrenta como principal barreira o processo de extração e aquisição desse conhecimento. Isso ocorre pois é difícil para especialistas explicarem quais são os atributos relevantes utilizados na inspeção visual e subsequente classificação. Dessa forma, é trabalhoso e difícil de obter regras explícitas para realizar a classificação (TIMMERMANS; HULZEBOSH, 1996).

Uma maneira de se contornar essa situação, que permite não somente uma melhor incorporação do conhecimento no sistema, mas também agrega maior flexibilidade para se lidar com diferentes variedades de produtos é a de se estabelecer um conjunto vasto de atributos a serem extraídos das imagens para realizar a classificação. Parte deles sendo estritamente relacionada com as informações coletadas junto aos especialistas enquanto outros são inferências e aproximações de atributos muito complexos ou computacionalmente ineficientes de serem obtidos. Nestes casos, os atributos devem ter uma alta correlação com os efetivamente utilizados na inspeção visual humana (por exemplo, pode-se utilizar a área do produto na imagem como uma aproximação do seu volume na cena).

Apesar de essa manobra auxiliar na incorporação do conhecimento tácito, um conjunto muito grande de atributos também irá introduzir informações redundantes, inconsistentes e até mesmo irrelevantes, prejudicando o processo de treinamento, pois o classificador terá que lidar com dados de alta dimensionalidade. Além disso, os algoritmos de aprendizado de máquina exploram a regularidade estatística dos dados fornecidos e, caso essa regularidade não ocorra entre as instâncias das classes, a taxa de erro dos classificadores escolhidos será elevada, independentemente da técnica de aprendizado adotada.

Felizmente, a literatura fornece alternativas para a remoção sistemática de atributos irrelevantes e redundantes por meio de um processo denominado de seleção de atributos (HALL, 1999). O problema de seleção de atributos envolve encontrar um bom subconjunto de atributos sob alguma função objetiva, por exemplo, a taxa de acerto das previsões do classificador (JOHN; KOHAVI; PFLEGER, 1994). Esse processo pode ser visto como uma tentativa de reduzir a dimensionalidade do espaço de dados utilizado no processo de reconhecimento de

padrões. Ao remover as características irrelevantes, a precisão de aprendizagem bem como a compreensão dos resultados tende a melhorar (DASH et al., 2002), aumentando assim a capacidade de generalização e interpretabilidade do modelo. Com o uso de técnicas de seleção de atributos, espera-se minimizar o problema criado pela necessidade de se trabalhar com um conjunto abrangente de atributos para garantir a incorporação de toda a informação relevante para o aprendizado da tarefa de categorização das mudas entre os diferentes níveis de qualidade. Apesar de focado na aplicação de mudas de violeta, os desafios aqui abordados se estendem para a classificação visual de diversos outros produtos agrícolas. Qualquer produto para os quais não existam regras formais para realizar sua distinção poderá se beneficiar dessa associação de técnicas.

1.2 **Objetivo**

O objetivo deste trabalho foi avaliar a influência de técnicas de seleção de atributos no desempenho de um sistema de visão computacional capaz de incorporar o conhecimento tácito de especialistas para realizar a classificação de mudas de plantas ornamentais entre diversos níveis de qualidade.

1.3 **Metodologia**

A primeira atividade necessária para a realização dos experimentos deste trabalho é a obtenção de um conjunto de imagens de mudas rotuladas por um especialista entre os diferentes níveis de qualidade. Esse conjunto rotulado de imagens serve de base para os desenvolvimentos necessários, desde o sistema de segmentação e extração de atributos das mudas como o sistema de classificação e validação dos resultados obtidos. A coleta das imagens é feita em ambiente com iluminação controlada, uma vez que variações no nível de iluminação podem alterar o resultado da segmentação de uma imagem para outra.

Os algoritmos de segmentação devem isolar regiões de *pixels* e encontrar pontos notáveis nas imagens que representem características que, segundo informações levantadas junto aos especialistas, sejam relevantes para a classificação das mudas entre os diferentes níveis de qualidade. É válido ressaltar

que um dos desafios do trabalho foi justamente estabelecer quais são as características mais relevantes para se realizar essa classificação.

Após validações iniciais, esses algoritmos poderão ser aprimorados para aumentar sua robustez perante variações de translação e rotação das mudas nas imagens, e também permitindo se extrair um conjunto maior de atributos, para se obter a melhor incorporação conhecimento, mencionada previamente.

Após o processamento, cada muda é representada por um conjunto de números que correspondem aos atributos obtidos por meio dos algoritmos de processamento de imagem. Dessa forma, o conjunto de mudas rotuladas é representado por um conjunto de dados alfanuméricos, que pode ser manipulado com maior facilidade e velocidade pelo sistema. É nesse conjunto de dados que serão aplicadas as diversas técnicas de seleção de atributos, visando reduzir a quantidade de atributos que serão utilizados nos classificadores à somente o subconjunto dos mais relevantes e menos ruidosos.

Para analisar o desempenho dos subconjuntos gerados, diferentes métodos de aprendizado de máquina (MITCHELL, 1997) serão utilizados. Uma vez que muitos das técnicas de seleção de atributos apenas ranqueiam o conjunto de atributos de acordo com sua relevância, uma busca exaustiva deverá ser realizada para determinar quantos atributos deverão ser incluídos no subconjunto final, a ser utilizado pelo classificador.

Essa busca será feita para cada uma das técnicas utilizadas e para cada classificador escolhido, gerando uma quantidade de testes igual a quantidade de técnicas de seleção vezes a quantidade de atributos vezes a quantidade de classificadores.

Inicialmente será considerado apenas o primeiro atributo mais relevante para a classificação, e esse subconjunto de atributos será fornecido para cada classificador. A taxa de acerto média será estimada através de validação cruzada *k-fold*. Nessa técnica, o conjunto total de instâncias é dividido em *k* grupos mutuamente exclusivos e de mesmo tamanho e, em seguida, *k-1* grupos são utilizados no treinamento dos classificadores. O grupo restante é então utilizado para estimar a acurácia. Esse processo é repetido *k* vezes, alternando entre os grupos formados. Dessa forma evita-se um viés no teste, pois a validação de cada classificador não considera instâncias de dados que foram utilizadas durante seu treinamento.

Ao final dos testes, se terá a variação das taxas de acerto, ou acurácia, de cada classificador para cada subconjunto de atributos possível, e também para cada técnica de filtragem. Desses resultados poder-se-á verificar qual classificador apresenta as melhores taxas de acerto e qual conjunto de atributos deverá ser utilizado para realizar a classificação das mudas de maneira mais otimizada.

1.4 Trabalhos correlatos

Existem na literatura diversos trabalhos que abordam problemas de inspeção visual de produtos agrícolas (tais como flores, frutas e outros produtos agrícolas), utilizando técnicas diversas. Brons et al. (1993) utilizaram Regressão Linear Múltipla, bem como Análise de Componentes Principais (PCA) para condensar a informação fornecida a uma Rede Neural (RN), no seu caso para classificar *Cyclamens* já crescidas. Foram utilizados quatro atributos: área foliar e área floral superior e lateral. Eles utilizaram duas câmeras, uma delas capturando imagens da vista superior dos vasos e outra da vista lateral.

Timmermans e Hulzebosch (1996) compararam o desempenho de análise estatística discriminante, utilizando abordagens linear e quadrática, contra a de RN para realizar segmentação por cor e também a classificação de violetas Africanas e Cactos. Foi realizada seleção de atributos, mas principalmente manualmente. No entanto, o valor *U-statistic* dos atributos obtidos foi considerado como uma medição do seu potencial discriminativo.

Ji e Singh (1996) exploraram as diferenças morfológicas nas imagens de plantas de Gerânio para segmentá-las e classificá-las, mas não mencionam o uso de técnicas de seleção de atributos. Foram utilizadas técnicas de limiarização para realizar a segmentação de imagens monocromáticas das plantas. Os autores obtiveram melhores resultados ao alterar uma técnica de limiarização adaptativa para esse fim.

Zheng; Sun e Zheng (2006), em uma revisão de técnicas de processamento de imagens, destacaram que o desempenho da classificação de produtos alimentares pode ser melhorada por meio da combinação de um maior número de características, como cor, tamanho, forma e textura, mas não mencionou como a seleção de atributos deve ser feita para determinar o conjunto exato que deve ser incluso. Parsons; Edmondson e Song (2009) estudaram a adição de medições

estereoscópicas na extração de características e utilizaram um modelo de regressão múltipla para investigar os efeitos desses atributos sobre os índices de qualidade de *Panys*, *Dianthus*, *Viola* e sobre o crescimento de plantas de *Ciclâmen* dentro de estufas, mas não foi feita seleção de atributos. Singh; Gupta e Gupta (2011) utilizaram momentos centrais como descritores de forma para classificar com sucesso diferentes espécies de bambus na Índia.

Blasco et al. (2009) relatam o desenvolvimento de uma máquina para classificar e separar arilos de romã, utilizando principalmente suas variações cromáticas. Eles detalham a estrutura mecânica da máquina, sua arquitetura computacional e os principais desafios enfrentados para viabilizar a incorporação do sistema, de forma a atender aos requisitos operacionais de uma produção de nível comercial. Os autores obtiveram resultados da ordem de 90% de acerto para a segmentação em tempo real dos arilos nas imagens, e conseguiram desenvolver com sucesso um protótipo que realizasse sua separação mecânica.

Unay et al. (2011) desenvolveram uma sistema de classificação de maçãs bicolores utilizando um sistema de visão multiespectral. Um total de 67 atributos foi extraído de cada maçã e uma etapa de seleção de atributos foi implementada para melhorar as taxas de acerto e reduzir o custo computacional da classificação das maçãs. Eles realizaram testes com cinco classificadores e obtiveram resultados até 9,2% melhores após a seleção de atributos.

Finalmente, Costa et al. (2011) fazem também uma revisão abrangente sobre o assunto, enfatizando a relevância das características morfológicas de produtos agrícolas para vários aspectos da avaliação de sua qualidade, incluindo triagem e classificação. Também foi proposta uma abordagem para se realizar a análise morfológica dos produtos de maneira automática, usando, como técnicas de seleção de atributos, a análise de componentes principais, a análise multivariada e a média de forma e variância (dentro de grupos de cultivares). Porém, o método de modelagem proposto se adéqua melhor a formas simples, como as de frutas, legumes e nozes e, portanto, não pode ser utilizado para a análise de formas mais complexas, como é o caso das mudas.

Pesquisas que abordem com efetividade as particularidades envolvidas na classificação e separação de mudas de flores e plantas ornamentais vêm se tornando mais importantes e não há trabalhos específicos sobre o assunto na

literatura. Além disso, a eficácia de técnicas de seleção de atributos em aumentar o desempenho de sistemas de visão computacional para controle de qualidade de produtos agrícolas não foi abordada nos trabalhos encontrados. Dessa forma, este trabalho visa propor, desenvolver e testar um sistema de classificação visual para atender às necessidades específicas da inspeção desse tipo de produto, avaliando o quanto o sistema é afetado por técnicas de seleção de atributos.

1.5 Organização do texto

Este trabalho está organizado da seguinte forma: no capítulo 2 apresenta o paradigma de sistemas de inspeção visual tradicionais e as principais dificuldades ao se tentar aplicá-los à indústria agrícola. Também discute as vantagens de se utilizar técnicas de aprendizado supervisionado de máquina para superar essas dificuldades e por fim os algoritmos utilizados no trabalho são apresentados.

O capítulo 3 discorre sobre o compromisso entre a utilização de um grande número de atributos para garantir que toda a informação necessária à classificação seja capturada e o desempenho dos classificadores frente a eventual presença de atributos irrelevantes e redundantes. Neste capítulo também se explica como técnicas de seleção de atributos podem ser usadas para maximizar esse compromisso e são apresentadas as principais abordagens utilizadas, assim como suas vantagens e desvantagens. Após justificar a abordagem adotada, enunciam-se as técnicas escolhidas para este trabalho.

O capítulo 4 detalha os procedimentos experimentais realizados para subsidiar a pesquisa, desde a coleta de dados até o procedimento de validação cruzada, utilizado para estimar as taxas de acerto obtidas nas diversas configurações avaliadas.

O capítulo 5 apresenta e discute os resultados observados após a realização dos experimentos, comentado cada etapa realizada: a filtragem de atributos, a geração dos subconjuntos e a comparação dos desempenhos dos classificadores utilizados.

Finalmente, o capítulo 6 apresenta as conclusões que puderam ser extraídas após realização do trabalho, assim como as limitações encontradas e sugestões de continuidade para a pesquisa.

2 SISTEMAS DE INSPEÇÃO VISUAL E CONTROLE DE QUALIDADE

Tradicionalmente, inspeções visuais e controle de qualidade são feitos por especialistas humanos (MITAL; GOVINDARAJU; SUBRAMANI, 1998). Apesar do fato de que em muitos casos os seres humanos podem realizar um melhor trabalho de inspeção do que as máquinas, os primeiros são mais lentos e se cansam rapidamente. Some-se a isso uma série de outras vantagens que sistemas artificiais apresentam em relação ao desempenho de especialistas humanos, como a dificuldade de se encontrá-los e treiná-los, a monotonia e repetitividade de alguns tipos de inspeção ou até mesmo sua complexidade e risco (e.g. indústria química, nuclear, etc.) e fica claro o motivo pelo qual sistemas de visão computacional têm sido cada vez mais empregados para automatizar esse tipo de tarefas nas mais diversas áreas.

2.1 Sistemas de Visão Computacional Industriais

Na maioria das indústrias em que esses sistemas são adotados, notam-se algumas características que os definem. A principal delas é a necessidade de uma padronização criteriosa do produto a ser analisado (GOLNABI; ASADPOUR, 2007).

Figura 1 - Dois exemplos de aplicação tradicional de sistemas de visão computacional na indústria.
Superior: bebidas engarrafadas. Inferior: peças metálicas.

Fonte: Curso de programação de software *In-Sight Explorer – Easy Builder*, Cognex Corporation - <http://www.omniimpex.com/workshops.html>

Nesses sistemas, a detecção de defeitos se baseia num padrão bem definido do produto, e variações mais acentuadas deste padrão são consideradas falhas. A decisão a ser feita pelo sistema é frequentemente binária, do tipo aceita ou rejeita, e as fronteiras para a tomada de decisão são bem delineadas (informação pessoal)¹. A Figura 1 fornece exemplos de duas aplicações típicas de sistemas de visão computacional industriais. No primeiro caso, das garrafas de bebidas, nota-se que a presença de manchas brancas na região da tampa caracteriza defeitos, e o sistema utiliza um limiar na quantidade de pixels brancos, determinado pelo operador do sistema, para tomar a decisão de aceite ou rejeite da garrafa. No segundo caso, das peças metálicas, há uma concatenação de critérios de decisão, que envolve a análise da presença de diversas regiões escuras na peça, manchas essas que representam os furos e chanfros que a peça deve conter para estar dentro do padrão aceite. Nota-se que, no exemplo, o sistema é capaz de lidar com a rotação da peça, requisito este que pode ser atendido através de técnicas de *template matching* (LEWIS, 1995). Ainda assim, o padrão do produto analisado é bem definido, cabendo ao sistema apenas compensar a rotação do objeto na cena antes de realizar as medidas das áreas escuras.

A decisão de aceite ou rejeite depende de todas as características analisadas. Basta que um dos furos ou que um dos chanfros não seja detectado para se rejeitar a peça. Uma tentativa de se sintetizar o funcionamento clássico de sistemas de visão computacional industriais pode ser vista na Figura 2.

2.1 Particularidades na Inspeção de Produtos Agrícolas

Vejamos agora outra situação, exemplificada na Figura 3. Nota-se que há diferenças importantes entre esse caso e os do exemplo anterior. A primeira e mais relevante é a dificuldade de se prontamente reconhecer qual o padrão definido para cada produto. Isso é especialmente verdadeiro para o caso de mudas agrícolas, que tem formações altamente complexas (HUANG; LEE, 2010). No caso das mudas de violeta, pode-se dizer que o tamanho das mudas tem influencia na decisão, porém fica claro que não é esta a única característica relevante para a diferenciação. A concentração de brotos na base das mudas parece ser importante também. Outros

¹ Curso de programação de software In-Sight Explorer – Easy Builder, Cognex Corporation. Realização Omni Sistemas de visão e IHM's, São Paulo, em 19 e 20 de setembro de 2012.

critérios como altura, largura, abertura também podem ter seu peso na diferenciação, porém não fica claro como considerar todos esses critérios simultaneamente na hora de tomar a decisão, como feito no exemplo anterior. Para as batatas, nota-se que o tamanho já não é de toda a importância e a presença de manchas e doenças parece ser relevante. Ainda assim, enfrenta-se o mesmo problema de como se considerar todos os critérios envolvidos para fazer a correta

Figura 2 - Diagrama das etapas do funcionamento de um sistema de visão computacional industrial. A tomada de decisão é feita a partir de regras explícitas fornecidas pelo usuário.

Fonte: Autor

classificação. Além disso, nota-se também uma segunda mudança no paradigma do exemplo anterior: a decisão deixa de ser binária, do tipo aceita ou rejeita. O que se tem nesse caso são níveis de qualidade decrescentes, ou classes de produtos.

Figura 3 - Exemplos de classificações de produtos agrícolas. Na parte superior: mudas de violeta por estaquia. Inferior: batatas lavadas.

Fonte: Autor

Tudo isso sugere que, no caso de produtos agrícolas, um sistema de visão computacional industrial tradicional não será a melhor solução para realizar a classificação dos produtos corretamente.

2.2 Aprendizado de Máquina

O campo de aprendizado de máquina, uma subárea da inteligência artificial, se dedica a tentar elaborar um modelo que seja capaz de prever uma ou mais variáveis de saída (categóricas ou quantitativas) tomando como base um conjunto de atributos que procuram descrever cada instância analisada. Para se construir esse modelo preditivo da variável de saída utiliza-se um conjunto de dados de treinamento.

Uma possível divisão para esse campo seria entre duas principais abordagens: o aprendizado de máquina supervisionado e o não supervisionado. No aprendizado de máquina supervisionado, a classificação é o problema de designar uma classe ou categoria para cada instância ou observação realizada. O classificador se baseia num conjunto de treinamento que contém diversos exemplos classificados corretamente, i.e., pares de instância e classe conhecidos. Ou seja, a variável de saída é parte integrante do conjunto de treinamento. Cada instância é

descrita por um número de atributos e o classificador deve mapear esse conjunto de atributos em uma única classe. Esse é um problema amplamente estudado e como resultado existem diversos algoritmos e abordagens para solucioná-lo (HALL, 1999; HASTIE, 2008).

No caso do aprendizado de máquina não supervisionado o conjunto de treinamento não conta com exemplos classificados. A divisão das instâncias entre diferentes classes é parte do problema em si. Em alguns casos é possível se determinar previamente qual o número de classes que se deseja obter, porém a determinação das classes é de responsabilidade do algoritmo de aprendizado em questão. O desafio é descrever como as instâncias se organizam ou ainda como melhor agrupá-las.

Nota-se que a junção de técnicas de aprendizado de máquina supervisionado com as de sistemas de visão computacional industriais permite lidar com relativa facilidade as questões levantadas no item 2.1. A dificuldade de se definir o padrão de cada produto é superada uma vez que, ao invés de se ter que fornecer regras explícitas sobre o padrão do produto, basta que se forneça um conjunto de exemplos corretamente classificados e o sistema será capaz de inferi-las automaticamente. Isso já levando em conta os todos os atributos que definem o produto (ou a instância) ao mesmo tempo. Além disso, a divisão em diversos níveis de qualidade (ou classes) ao invés da decisão binária faz parte do funcionamento interno dos classificadores, que não raro podem trabalhar com qualquer número de classes que seja necessário.

Desta forma, podemos redesenhar o diagrama da Figura 2 para representar sistemas de visão computacional que incorporam técnicas aprendizado de máquina supervisionado para realizar a inspeção. Esses sistemas são os mais aptos a lidar com a variabilidade e peculiaridades aqui discutidas, envolvidas na classificação de produtos agrícolas, mas podem e são também aplicados na indústria tradicional que, em casos específicos, pode também se beneficiar da flexibilidade adicional que apresentam (MALAMAS et al., 2003). Uma representação gráfica desse sistema pode ser vista na Figura 4.

Figura 4 – Diagrama das etapas do funcionamento de um sistema de visão computacional que utiliza aprendizado de máquina supervisionado ao realizar a inspeção. O sistema precisa ser previamente treinado com exemplos para inferir as regras de classificação.

Fonte: Autor

2.3 Algoritmos de classificação

A ênfase dada à seleção de atributos para se realizar a classificação pode ser usada para diferenciar os diversos algoritmos de aprendizado. De um lado estão os que tendem a utilizar todos os atributos fornecidos, e, portanto toda a informação disponível (e.g. *nearest neighbors*). Na outra ponta tem-se, por exemplo, as árvores de decisão, que dão grande peso aos atributos mais importantes e tendem a ignorar os menos relevantes (HALL; HOLMES, 2003). De maneira a subsidiar as pesquisas, foram selecionados para esse trabalho diversos algoritmos amplamente utilizados na literatura, procurando abranger o espectro de possibilidades quanto a ênfase dada à seleção de atributos. Além disso, outro fator que justificou a escolha desses algoritmos em detrimento de outros foi o fato de todos estarem implementados no pacote de software utilizado na realização dos experimentos do trabalho. São eles: *k-Nearest Neighbors*, *Naïve Bayes*, *Logistic Regression*, *C4.5*, *Multilayer Perceptron* e *Random Forest*. Uma breve descrição é fornecida a seguir.

2.3.1 k-Nearest Neighbors (kNN)

Este é um dos mais simples algoritmos utilizados em aprendizado de máquina, cuja regra foi introduzida por Cover e Hart (1967). Ele classifica as observações de acordo com a classe mais comum dentre seus k vizinhos mais próximos no espaço de características formado (de acordo com alguma métrica de distância – usualmente a distância Euclidiana). No entanto, a presença de atributos ruidosos ou até mesmo irrelevantes pode reduzir drasticamente a acurácia da classificação. Além disso, este algoritmo apresenta a desvantagem de ser não paramétrico, necessitando, portanto de todo o conjunto de dados disponíveis para tomar sua decisão. A Figura 5 ilustra a abordagem deste algoritmo para o problema de classificação. No exemplo, o classificador é um 3NN, pois avalia os três vizinhos mais próximos da instância para tomar a decisão. A ocorrência de atributos irrelevantes ou inconsistentes tende a rapidamente reduzir as taxas de acerto do classificador gerado, pois o espalhamento das amostras no espaço dimensional analisado será cada vez maior e com mais sobreposições.

Figura 5 – Representação gráfica do processo de decisão do kNN.

Fonte: Autor

2.3.2 Naïve Bayes

O *Naïve Bayes* é um dos classificadores mais simples e difundidos na literatura. Como abordado por Rish (2001), apesar da forte suposição teórica que carrega, de que os atributos são independentes entre si, pode apresentar desempenhos comparáveis aos outros algoritmos mais complexos. Uma de suas vantagens é que ele não apresenta parâmetros de configuração, tornando sua utilização ainda mais simples e imediata. Por meio da simples contagem das amostras de treino, ele estima as probabilidades $p(C)$ de ocorrência de cada classe

c e as probabilidades condicionais $p(A_i|C)$ de ocorrência de cada atributo a_i com relação a cada classe c . Assim, dada uma observação de todos os n atributos $a_1, \dots, a_n$, a classe R estimada pelo classificador advém de uma regra de decisão de *maximum a posteriori*, i.e.:

$$R(a_1, \dots, a_n) = \underset{c}{\operatorname{argmax}} p(C = c) \prod_{i=1}^n p(A_i = a_i | C = c) \quad (1)$$

2.3.3 Logistic Regression

Este classificador realiza uma análise de regressão no conjunto de treinamento, de forma análoga a realizada no método dos mínimos quadrados (JORDAN, 2002). Porém, ao invés de utilizar uma combinação linear para atribuir pesos aos atributos, o *logistic regression* utiliza a função sigmoidal $g(s) = \frac{1}{1+e^{-s}}$, também chamada de função logística, justificando assim o seu nome. A Figura 6 ilustra o comportamento dessa função não linear.

Figura 6 – Função logística.

Fonte: Autor

No *logistic regression*, para o caso de um problema de classificação binário, toma-se s como uma combinação linear dos atributos X_i ponderados pelos parâmetros w_i ,

$$s = w_0 + \sum_{i=1}^n w_i X_i \quad (2)$$

gerando em seguida as seguintes probabilidades condicionais:

$$P(Y = 1|s) = \frac{1}{1 + e^{-s}} \quad (3)$$

$$P(Y = 0|s) = 1 - P(Y = 1|s) \quad (4)$$

Onde $P(Y = 1|s)$ é a probabilidade da instância representada por s pertencer a classe 1 e $P(Y = 0|s)$ a de pertencer a classe 0. O processo de treinamento envolve ajustar os valores de w_i de forma a maximizar a taxa de acerto do classificador através da estimativa por máxima verossimilhança (HALLORAN, 2009). A decisão sobre a qual classe uma nova instância pertence é feita analisando-se a razão entre as probabilidades das equações (3) e (4), na forma $\frac{P(Y=0|s)}{P(Y=1|s)}$. Caso o valor seja maior que a unidade, a instância será atribuída à classe 0, caso contrário à classe 1.

Problemas com mais do que duas classes são resolvidos por meio de técnicas de generalização estatísticas, que avaliam a pertinência de uma instância a uma classe caso a caso, similarmente ao *Naïve Bayes*. Ou seja, estima-se a probabilidade para cada classe e a escolhida é a que apresentar o maior valor.

2.3.4 Multilayer Perceptron

Redes Neurais Artificiais surgiram na última década e tem sido muito utilizadas para resolver problemas complexos do mundo real de diversas áreas diferentes, indo desde biologia até finanças. A arquitetura mais popular para as redes neurais é a do *Multilayer Perceptron* (MLP), com seu algoritmo de treinamento de *Backpropagation* (RUMELHART; HINTONT; WILLIAMS, 1986, BASHEER; HAJMEER, 2000). As vantagens das redes neurais nessa configuração são a sua capacidade de distinguir dados não lineares, possibilidade de paralelismo e o fato de ser comprovadamente capaz de aproximar arbitrariamente bem qualquer função contínua de números reais (CYBENKO, 1989). O MLP utiliza funções de ativação sigmóides e caracteriza-se por apresentar múltiplas camadas de nós num grafo orientado, sendo cada camada totalmente conectada a próxima, como pode ser visto no exemplo da Figura 7.

Figura 7 – Esquema de grafo orientado com camadas totalmente conectadas.

Fonte: Autor

2.3.5 C4.5

O C4.5 cria uma árvore de decisão utilizando o conjunto de treino, ele escolhe a cada nó da árvores o atributo que a melhor divide em outros dois subconjuntos. A Figura 8 exemplifica o conceito de uma árvore de decisão que, no caso ilustrado, faz verificações tanto discretas quanto contínuas para definir se um indivíduo terá sucesso financeiro na sua vida profissional.

Figura 8 – Exemplo de árvore de decisão sobre sucesso na carreira.

Fonte: Autor

O critério usado para escolher o atributo de divisão é o ganho de informação normalizado que ele apresenta, i.e., a diferença de entropia resultante de uma divisão feita com ele (QUINLAN, 1993). O C4.5 é uma extensão do algoritmo ID3, também proposto por Quinlan, que realiza algumas melhorias no seu predecessor.

As principais delas são a capacidade de lidar com valores contínuos e discretos e a de processar conjuntos de treinamento com valores de atributos faltantes.

2.3.6 Random Forest

O *Random Forest* (RF) (BREIMAN, 2001) é um classificador que combina diversas árvores de decisão usando a classe eleita com maior frequência como saída. Ele é tido como um dos algoritmos mais precisos, tendo sido testado em diversas bases de dados (CARUANA; NICULESCU-MIZIL, 2006). Além disso, ele tem a vantagem de se adequar para a manipulação de grandes conjuntos de dados, dando maior ênfase a atributos mais relevantes e menor ênfase para os menos relevantes no momento de realizar a classificação, conforme discutido em (HALL; HOLMES, 2003). Seu pseudocódigo é representado no Algoritmo 1.

Algoritmo 1: Pseudocódigo para o Random Forest

```

 $n_c$ : número de classificadores a serem construídos;
 $x$ : número de atributos para cada árvore;
 $f$ : número total de atributos;

for  $i = 1$  to  $n_c$  do
  Amostrar aleatoriamente o conjunto de treino  $D$  com
  substituição para produzir  $D_i$ ;
  Criar um nó raiz,  $N_i$  contendo  $D_i$ ;
  Chamar ConstruirArvore( $N_i$ );
end for

ConstruirArvore( $N$ ):
  if  $N$  contém instâncias de apenas uma classe then
    return;
  else
    Selecionar aleatoriamente  $x$  atributos possíveis para
    particionar  $N$ ;
    Selecionar o atributo  $F$  com maior ganho de informação ao
    particionar;
    Criar  $f$  nós filhos de  $N$ ,  $N_1, \dots, N_f$ , onde  $F$  tem  $f$  valores
    possíveis ( $F_1, \dots, F_f$ );

    for  $i = 1$  to  $f$  do
      Preencher  $N_i$  até  $D_i$ , onde  $D_i$  são todas as instâncias de  $N$ 
      coincidentes com  $F_i$ ;
      Chamar ConstruirArvore( $N_i$ );
    end for
  end if
end ConstruirArvore

```


2.4 Considerações finais do capítulo

Neste capítulo mostrou-se o funcionamento básico de sistemas de visão tradicionais, analisando-se suas capacidades e limitações quando aplicados para a inspeção e classificação de produtos agrícolas. As particularidades dessa categoria de produtos apresentam desafios que dificilmente podem ser solucionados puramente com tecnologias de visão computacional. A dificuldade de se explicitar regras para a classificação, assim como a de se considerar diversos fatores simultaneamente para a tomada de decisão inviabilizam essa opção.

Viu-se que técnicas de aprendizado de máquina supervisionado são a solução ideal para esse problema, pois quando associadas a sistemas de visão computacional, permitem lidar com facilidade com as dificuldades apresentadas.

Frente a diversidade de algoritmos disponíveis e a fim de subsidiar as pesquisas, alguns deles foram escolhidos e detalhados neste capítulo. Procurou-se eleger algoritmos que abrangessem o espectro de alternativas referentes a ênfase que dão a seleção de atributos.

Os próximos capítulos tratarão de aprofundar o entendimento sobre as técnicas de seleção de atributos utilizadas e de então detalhar o experimento realizado e os resultados obtidos.

3 SELEÇÃO DE ATRIBUTOS

Muitos fatores afetam o sucesso do aprendizado de máquina nas suas diversas aplicações. Dentre eles, a representação e a qualidade dos dados utilizados são alguns dos mais importantes (HALL, 1999). No campo de aprendizado de máquina, a seleção de atributos consiste no processo de determinação de um subconjunto que contenha apenas atributos relevantes para o aprendizado da tarefa.

O que justifica a adoção da seleção de atributos em uma dada aplicação é a suposição de que o conjunto de dados utilizado contenha atributos redundantes ou irrelevantes para a tarefa a ser aprendida. Existem também outros benefícios em potencial ao se utilizar a seleção de atributos: facilitar a visualização e compreensão dos dados, diminuir as necessidades de armazenamento e manipulação dos dados e também reduzir o tempo de treinamento e execução dos classificadores (GUYON; ELISSEEFF, 2003).

Isto posto, uma análise na literatura permite entender que irrelevante é um atributo que, independentemente do contexto, não fornece nenhuma informação útil para a classificação. Em outras palavras, um atributo irrelevante é o que não é relevante. Existem porém diversas definições para a relevância de um atributo. Uma síntese destas definições pode ser vista em John; Kohavi e Pfleger (1994), que inclusive dividem a relevância dos atributos entre forte e fraca. A noção de redundância é também bastante discutida na literatura. No consenso geral, um atributo é dito redundante quando não acrescenta nenhuma informação no sistema além da já disponibilizada por outros atributos. As formalizações de redundância são expressas principalmente em termos da correlação entre os atributos utilizados, como pode ser visto em (HALL, 1999). Existem porém abordagens diferentes, como a apresentada por Yu e Liu (2004), que utilizam o conceito de *Markov blanket* para definir redundância. Essa definição se contrapõe de alguma forma a suposição de que atributos altamente correlacionados são sempre redundantes, por não acrescentarem nova informação ao sistema, o que nem sempre é verdadeiro, como demonstrado por Guyon e Elisseeff (2003). Isto posto, vale ressaltar que o foco deste trabalho é o de avaliar se seleção de atributos, independentemente da definição formal escolhida, apresenta vantagens na classificação de mudas de plantas ornamentais, na qual a obtenção dos atributos é feita de maneira empírica e

dos quais a relevância para a classificação é incerta. Premissa essa que, conforme mencionado no início desta seção, justifica a utilização de técnicas de seleção de atributos.

3.1 Abordagens para seleção de atributos

Técnicas de seleção de atributos podem ser amplamente divididas entre duas categorias: filtros e *wrappers* (invólucros). Existe também uma terceira categoria, porém menos intercambiável do que as duas primeiras, chamada de *embedded*.

3.1.1 Filtros

Filtros avaliam a relevância dos atributos examinando apenas as características intrínsecas ao conjunto de dados disponível, eles podem ser entendidos como uma etapa de pré-processamento dos dados (KOHAVI; JOHN, 1997). Isso pode ser feito de duas maneiras: avaliando o mérito de cada atributo individualmente, de acordo com alguma métrica, gerando uma lista ordenada de atributos de acordo com a nota que recebem, sendo o primeiro o mais relevante e o último o menos relevante. Normalmente os atributos com notas muito baixas são removidos e o subconjunto restante é apresentado ao classificador. Essa abordagem é chamada de *feature ranking*.

Outra maneira de realizar seleção de atributos através de filtragem é avaliar o mérito de conjuntos de atributos ao invés de considerar cada atributo individualmente. É o que se chama de seleção de subconjunto (*subset selection*). Neste caso, a lógica é a similar a utilizada no *feature ranking*, porém o que se faz é a classificação de subconjuntos candidatos. Nesse caso, a métrica utilizada deve ser capaz de considerar diversos atributos simultaneamente e uma busca também deve ser associada ao método para tornar viável percorrer o espaço possível de subconjuntos.

3.1.1.1 Vantagens e desvantagens

As principais vantagens de técnicas de filtragem é que elas podem ser usadas em conjuntos de dados muito grandes, pois apresentam maior escalabilidade. São

computacionalmente simples e rápidas e independentes do algoritmo de classificação escolhido. Ou seja, a seleção pode ser realizada apenas uma vez, e avaliada por diversos classificadores. A principal desvantagem dos filtros é que eles completamente ignoram a futura interação com o classificador. Além disso, a maioria das técnicas propostas são do tipo *feature ranking*, que ignora também eventuais dependências entre os atributos (KOHAVI; JOHN, 1997, SAYES; INZA; LARRAÑAGA, 2007).

3.1.2 Wrappers

Wrappers realizam a seleção de atributos sempre na abordagem de seleção de subconjunto, ou seja, comparam sempre conjuntos de atributos. A seleção é feita utilizando um algoritmo de classificação como uma caixa preta, sem necessitar de nenhuma informação sobre o algoritmo em si, preocupando-se apenas com a sua interface. Um esquema de busca é associado ao *wrapper*, pois o espaço de subconjuntos a serem considerados cresce exponencialmente com o número de atributos. A busca utiliza o algoritmo de classificação como parte da função de avaliação do mérito de cada subconjunto. A avaliação é feita treinando e testando o classificador, de acordo com o algoritmo escolhido. Ou seja, a busca no espaço de subconjuntos possíveis “envolve” ou “embrulha” o classificador. A suposição é a de que o algoritmo de classificação que irá futuramente utilizar o subconjunto de atributos gerados para realizar a classificação deve fornecer uma melhor estimativa sobre o mérito de cada subconjunto do que uma outra métrica qualquer (LANGLEY, 1994).

3.1.2.1 Vantagens e desvantagens

A principal vantagem dos *wrappers* é o fato deles garantirem a interação entre a busca pelo melhor subconjunto e o algoritmo de classificação que será utilizado. Isso garante também que sejam levadas em conta as dependências entre os atributos. Além disso, muitas vezes se obtém melhores resultados com *wrappers* do que com filtros, principalmente por avaliarem o melhor subconjunto para o classificador específico que está sendo utilizado. Por outro lado, eles demoram muito mais do que filtros para executar, pois precisam repetidamente treinar e testar

classificadores para avaliar cada novo subconjunto. Além disso, todo o processo precisa ser repetido quando o algoritmo de classificação for substituído. Finalmente, *wrappers* também apresentam um risco maior de hiperespecialização do que os filtros (HALL, 1999, HALL; HOLMES, 2003, SAYES; INZA; LARRAÑAGA, 2007).

3.1.3 *Embebbbed*

Técnicas de seleção de atributos *embebbbed* realizam a busca por um subconjunto ideal de atributos enquanto treinam o classificador. Ou seja, a seleção de atributos é parte inerente do processo construtivo do classificador. De forma análoga aos *wrappers*, essa abordagem também é sintonizada para cada algoritmo de aprendizado, que deve obrigatoriamente contar com funcionalidades de seleção de atributos por projeto. Alguns exemplos são as próprias árvores de decisão, que podem não considerar alguns atributos no classificador gerado ou até mesmo algoritmos que atribuem coeficientes cada vez menores a atributos que se mostram irrelevantes durante o treinamento, chegando a zerá-los (e, conseqüentemente, removê-los).

3.1.3.1 *Vantagens e desvantagens*

Assim como nos *wrappers*, a seleção de atributos “embarcada” tem a vantagem de considerar a interação com o algoritmo de aprendizado utilizado, com a vantagem adicional de apresentarem um custo computacional muito mais reduzido (SAYES; INZA; LARRANAGA, 2007). A principal desvantagem é que não podem ser utilizados em parceria com outros algoritmos de aprendizado, pois a seleção é feita internamente, durante o processo construtivo.

3.2 Abordagem escolhida no trabalho

O objetivo do trabalho é o de verificar a eficácia de técnicas de seleção de atributos na classificação de mudas de flores ornamentais. A forma proposta para realizar essa avaliação é realizar uma comparação das taxas de acerto obtidas por diversos classificadores ao utilizar os subconjuntos gerados por essas técnicas.

Os resultados dos filtros independem do algoritmo de classificação a ser utilizado, eles podem ser fornecidos a diversos classificadores, garantindo isonomia entre as comparações das taxas de acerto obtidos. Além disso, os filtros têm a vantagem de executarem diversas vezes mais rápido do que os *wrappers*, que tem custo computacional proibitivo e podem se tornar inviáveis para grandes conjuntos de dados (HALL, 1999). Não obstante, técnicas embarcadas são intrínsecas à algoritmos de aprendizado que, por projeto, realizam a seleção de atributos durante a construção do classificador, inviabilizando sua utilização com outros classificadores e, consequentemente, inviabilizando a comparação de resultados.

Sendo assim, neste trabalho optou-se por utilizar apenas técnicas de filtragem para realizar a seleção de atributos. Uma descrição de cada uma delas é fornecida na seção a seguir.

3.3 Técnicas de seleção de atributos

Foram escolhidas algumas das técnicas de filtragem mais populares na literatura e que também estivessem implementadas nas bibliotecas do software utilizado no experimento deste trabalho, o Weka (WITTEN et al., 1999), desenvolvido pela Universidade de Waikato, na Nova Zelândia. As explicações sobre o funcionamento de cada um dos métodos escolhidos são apresentadas a seguir.

3.3.1 Correlation-based Feature Selector (CFS)

CFS é um dos métodos de filtragem utilizado neste trabalho. Ele trabalha por seleção de subconjunto, ou seja, avalia diversos atributos simultaneamente ao invés de avaliá-los individualmente. Este método aplica uma função heurística baseada em correlação para avaliar e classificar o espaço de subconjuntos de atributos, usando uma busca *best-first*. O objetivo da avaliação é obter um subconjunto em que os atributos são altamente correlacionadas com uma classe específica, enquanto não correlacionados com nenhum dos outros atributos. Essa avaliação está de acordo com Hall (1999) e sua definição de um subconjunto relevante: "Um subconjunto bom é o que contém atributos altamente correlacionados com a

(previsão da) classe, mas não correlacionados (não preditivos de) uns com os outros".

Neste método, atributos irrelevantes tendem a ser ignorados uma vez que a sua correlação com a classe será baixa. Atributos redundantes ficam também destinados à serem descartados, uma vez que estão altamente correlacionados com um ou mais dos outros atributos. A função de avaliação usada é mostrado na Equação (5) abaixo:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (5)$$

Onde k é o número de atributos em um dado subconjunto S , $\overline{r_{cf}}$ é a média da correlação atributo-classe, $\overline{r_{ff}}$ é a intercorrelação média entre atributo-atributo e M_s representa o mérito de S . Assim que cinco iterações consecutivas não apresentarem melhor avaliação do subconjunto anterior, a busca é finalizada.

3.3.2 Information Gain

Information Gain (Info Gain) é outro filtro utilizado para realizar a seleção de atributos. Neste caso porém, a abordagem do filtro é a de *feature ranking*, que avalia o mérito de cada atributo individualmente. Como o nome sugere, ele considera quanta informação é obtida para a classificação, dado que o atributo seja considerado. Ele mede a variação de entropia na informação, dado um atributo F_i de acordo com a Equação (6):

$$InfoGain(C, F_i) = H(C) - H(C | F_i) \quad (6)$$

Onde $H(C)$ é a entropia da classe C e $H(C | F_i)$ é a entropia da classe C , dado F_i .

Usando essa métrica, os atributos podem ser ordenados de acordo com os seus resultados individuais, dependendo da quantidade de informação que fornecem para a classificação.

3.3.3 Gain Ratio

Gain Ratio é muito semelhante ao *Information Gain*. Como pode ser visto a partir da Equação (7), a principal diferença dessa métrica em relação à anterior é que ela mede o ganho na informação para a classificação *em relação* à entropia de um dado atributo F_i :

$$GainRatio(C, F_i) = \frac{H(C) - H(C | F_i)}{H(F_i)} \quad (7)$$

Onde $H(C)$ é a entropia da classe C , $H(C|F_i)$ é a entropia da classe C dado F_i e $H(F_i)$ é a entropia do atributo F_i .

3.3.4 Chi-Quadrado

Chi-quadrado é um teste estatístico muito comum que mede a independência de um atributo para a classe. Sendo assim, pode ser prontamente utilizado como uma métrica para a seleção de atributos por *feature ranking*. Dada a hipótese de que um atributo é independente de uma classe, a divergência entre as distribuições Chi-quadradas observadas e esperadas é calculada para testá-la. No caso, as características mais relevantes são as que mais rejeitam essa hipótese. Então, ao fazer isso para todos os atributos, eles podem ser classificados de acordo com os valores obtidos. A fórmula do Chi-Quadrado é a seguinte:

$$\chi^2 = \sum \frac{(F_o - F_e)^2}{F_e} \quad (8)$$

Onde F_o é a frequência observada e F_e é a frequência esperada.

3.4 Considerações finais do capítulo

Neste capítulo mostrou-se as principais abordagens existentes para seleção de atributos, assim como suas principais vantagens e desvantagens. Também se viu uma breve síntese e revisão bibliográfica das definições de relevância, irrelevância e redundância. A natureza do objetivo e da metodologia propostas para a realização do trabalho fizeram com que se optasse pela utilização de técnicas de filtragem para realizar a seleção de atributos previamente à classificação das mudas. Sendo assim,

no final do capítulo temos uma descrição detalhada de cada método de filtragem adotado, todos já implementados na biblioteca de algoritmos Weka e amplamente difundidos na literatura.

4 PROCEDIMENTOS EXPERIMENTAIS

Para se atingir o objetivo proposto para o trabalho concebeu-se um conjunto de etapas experimentais com o intuito de se gerar o material que viabilizasse as comparações e avaliações propostas na metodologia, item 1.3. As seções a seguir detalham cada uma dessas etapas.

4.1 Coleta de dados

O experimento foi realizado com mudas de violeta africana (*Saintpaulia*), geradas a partir do estaqueamento de plantas matrizes já crescidas, que é a técnica tradicionalmente utilizada para a produção comercial de flores e plantas ornamentais. Todo o material foi fornecido por uma estufa comercial localizada na cidade de Holambra, no interior do estado de São Paulo. As mudas foram coletas durante a primavera de 2011.

O rotulamento das mudas, que se constitui na determinação de a qual classe pertence cada uma delas, foi feito pelo proprietário do sítio, que é um especialista local com mais de 20 anos de experiência no cultivo de violetas. As mudas foram classificadas em quatro níveis de qualidade decrescente, de acordo com seu potencial de crescimento. As mudas de melhor qualidade foram atribuídas a classe A, as de segunda melhor qualidade a classe B, as de terceira qualidade a classe C e finalmente as de pior qualidade atribuídas a classe D. Esse foi o maior número de classes com o qual o especialista se sentiu confortável para realizar a classificação com consistência. A Figura 9 mostra um exemplo das mudas rotuladas em cada classe. Foram obtidas aproximadamente 75 mudas para cada classe, totalizando quase 300 mudas classificadas para a realização do experimento. Esse valor foi o máximo de mudas por classe obtidos dentro da disponibilidade de tempo do especialista para realizar a classificação.

A captura das imagens foi feita em ambiente controlado para evitar diferenças entre as imagens. As mudas foram colocadas e retiradas do anteparo manualmente. Um esquema do ambiente de captura pode ser visto na Figura 10. Foram utilizadas para a iluminação duas lâmpadas tubulares Philips, modelo TL-D 15W/75-650. Um domo difusor semicilíndrico foi utilizado para minimizar a ocorrência de sombras e reflexões especulares (manchas brancas devido ao excesso de iluminação).

Figura 9 – Exemplos de imagens das mudas classificadas nos diversos níveis em ordem decrescente de qualidade, sendo (a) um exemplo de muda da melhor qualidade e (d) da pior.

Fonte: Autor

A câmera utilizada para conduzir a aquisição das imagens foi a Logitech QuickCam Pro 4000, com resolução de 1.3 *Megapixels*, colorida. O balanço de branco automático da câmera foi desabilitado para evitar flutuações de cores entre as imagens. O tamanho das imagens foi especificado para 640 x 480 *pixels* e a distância entre as lentes da câmera e o anteparo das mudas foi de 20 centímetros. Nessa configuração, a área de varredura da câmera era de 165 x 123 milímetros. A superfície do anteparo utilizado era da cor azul, para maximizar o contraste entre objeto e fundo.

Foram obtidas duas imagens de cada muda, ambas com a base da raiz para o lado direito da imagem, por motivos que serão vistos adiante. Para fazer a segunda imagem, a muda era rotacionada em 180 graus em relação ao seu eixo de simetria longitudinal, simulando-se assim duas possíveis posições que a muda poderia assumir no caso de estarem sendo colocadas numa esteira motorizada para a automação da classificação. Dessa forma foram geradas 607 imagens de mudas rotuladas, cerca de 150 de cada classe. Um arquivo com todas as imagens está disponível gratuitamente na rede, através do endereço: <https://sites.google.com/site/seedlingsdatabase/home>.

Figura 10 – Esquema do ambiente de captura utilizado para obter as imagens das mudas.

Fonte: Autor

4.2 Binarização

A binarização das imagens consiste na determinação de quais regiões da imagem pertencem ao fundo e quais pertencem ao objeto de interesse (TRIER; JAIN, 1995). A primeira abordagem utilizada neste trabalho baseou-se no método proposto por Blasco et al. (2009), feito por meio da análise individual de cada pixel, utilizando uma limiarização no espaço de cor RGB, de acordo com características do histograma de regiões da imagem pertencentes ao objeto em oposição às características apresentadas pelo histograma de regiões do fundo. Esse método de binarização foi implementado inicialmente, e diversas configurações de limites foram experimentadas, porém nenhuma delas foi adequada para o sistema. Essa técnica se mostrou altamente sensível a variações nas condições de iluminação e pouco robusta na eliminação de objetos alheios, como partículas de terra e areia, trazidos junto com as mudas.

Dessa forma, uma segunda abordagem para a binarização das imagens foi adotada, dessa vez inspirada em técnicas de *chromakeying* e representada matematicamente como:

$$T_{\alpha,\beta,\gamma,k}(R, G, B) = \{x \in E \mid \alpha \cdot R(x) + \beta \cdot G(x) + \gamma \cdot B(x) > k\}, \quad (9)$$

onde T é o operador de limiarização de imagens coloridas, x representa cada pixel da imagem, E representa o objeto de interesse, α , β , γ e k são os parâmetros de

limiarização do operador, e $R(x)$, $G(x)$ e $B(x)$ são os valores de intensidade do pixel x nos canais vermelho, verde e azul, respectivamente. Os valores foram ajustados manualmente, chegando-se na configuração de $\alpha = 2$, $\beta = 4$, $\gamma = -4$ e $k = 0$, adotada para a binarização das mudas. O resultado desse método de binarização se mostrou robusto quando aplicado a toda a coleção de 600 imagens das mudas. Um exemplo de sua aplicação pode ser visto na Figura 11b.

Figura 11 - (a) Imagem original. (b) Resultado intermediário, após a binarização. (c) Resultado final, após aplicação de filtros mediana e abertura morfológica.

Fonte: Autor

4.3 Extração de atributos

Os atributos extraídos das imagens das mudas foram estabelecidos a partir de informações fornecidas pelo especialista sobre quais características das mudas ele considerava importantes para realizar a separação. Além disso, parte dos atributos também foram estabelecidos arbitrariamente para garantir a máxima captura de informação para o sistema de classificação. As principais características mencionadas pelo especialista foram o volume das mudas, a ocorrência das folhas aos pares ao invés de isoladamente e também a concentração de brotos na base da muda.

A remoção de resquícios de raízes e outras partículas foi feita aplicando-se um filtro mediana, no qual uma janela é centrada no pixel de interesse e os valores

das intensidades de cada *pixel* pertencente à janela são ordenados. A mediana desses valores é calculada e é estabelecida como o novo valor do pixel central, sendo um filtro muito útil na remoção de ruído do tipo “sal e pimenta” (WANG; ZHANG, 1999). Em seguida é realizada uma operação morfológica de abertura na imagem (erosão seguida de dilatação) e uma segunda aplicação do filtro mediana.

A resolução espacial de cada pixel foi calculada por meio da média do comprimento em milímetros da imagem dividido pelo número de colunas da imagem e da sua largura dividida pelo número de linhas. Dessa forma, assimetrias no tamanho real dos *pixels* na cena são aproximadas pelo seu valor médio de tamanho, minimizando erros de estimativa do tamanho real dos objetos na cena.

O primeiro atributo calculado é a área total da muda, representada pela soma de todos os *pixels* de interesse (brancos) da imagem, convertida em seguida para milímetros quadrados utilizando essa resolução espacial.

Para manter a coerência na extração de diversos atributos das mudas, como seu comprimento e largura, as medições devem ser feitas em relação ao seu eixo de simetria longitudinal. A obtenção desse eixo pode ser feita unindo-se o centroide da muda com a base da sua raiz. O cálculo e a determinação do centroide podem ser feitos sem dificuldades, utilizando-se a média das coordenadas dos pontos da imagem que pertencem à muda. Em contrapartida, o ponto que representa a base da raiz não conta com um método eficaz para ser determinado. Impôs-se então, a limitação de que todas as mudas deveriam estar com a base da raiz direcionada para o lado direito da imagem e, com essa restrição, esse ponto foi obtido calculando-se o centroide de uma sub-região da imagem de 10 x 20 mm, alinhada à direita e centrada verticalmente no ponto mais à direita da imagem pertencente a muda. Essa sub-região foi utilizada para corrigir eventuais distorções da base da muda, que poderiam deslocar o eixo de simetria significativamente. Um exemplo da obtenção desse eixo pode ser visto na Figura 12.

Em seguida é calculado o menor retângulo que envolve a muda, alinhado ao eixo. Dele se extraem cinco atributos: comprimento, largura e alongação (razão entre comprimento e largura) da muda, assim como a área total do retângulo envolvente e a ocupação do retângulo, i.e. a razão entre a área de muda sobre a área do retângulo.

O menor círculo envolvente é também calculado, gerando outros três atributos: área do círculo, ocupação do círculo e a distância entre seu centro e o centroide da muda. A distância entre a base da raiz (ponto B da Figura 12) e o centroide da muda (ponto C da Figura 12) é também computada, como uma maneira de estimar a concentração de brotos na base da muda.

Finalmente, a distribuição de massa da muda é calculada por meio de 16 sub-regiões, que podem ser vistos na Figura 13. Cada uma delas representa 1/16 da área total do retângulo envolvente. As taxas de ocupação de cada uma dessas subáreas são também utilizadas como atributos.

Figura 12 - Exemplo da obtenção do eixo de simetria longitudinal da muda. O ponto A é o na extrema direita da imagem que pertence a muda, B é o centroide da região destacada e C representa o centroide da muda. A união de B com C estabelece o eixo desejado.

Fonte: Autor

Utilizou-se a biblioteca OpenCV 2.3.1 (BRADSKI; KAEHLER, 2008) para realizar todo o processamento de imagem envolvido neste trabalho. Para homogeneizar a importância dos atributos ao se realizar a classificação, todos foram normalizados para a faixa de valores reais [0, 1].

Em suma, são extraídos um total de 26 atributos das imagens das mudas, listados em conjunto a seguir: área da muda, comprimento, largura, alongação, área do retângulo envolvente, ocupação do retângulo, área do círculo, ocupação do

círculo, distância centroide-círculo, distância base-centroide, além das ocupações das 16 sub-regiões.

Figura 13 – Visualização dos atributos extraídos das imagens. Os retângulos numerados representam sub-regiões que avaliam a distribuição de massa das mudas.

Fonte: Autor

4.4 Definindo os subconjuntos

Information Gain, Chi-quadrado e Gain Ratio são, como discutido previamente, métodos de filtragem que realizam o *feature ranking*, avaliando individualmente os atributos disponíveis de acordo com sua relevância para o aprendizado da tarefa. Em outras palavras, essas técnicas medem o poder discriminativo de cada atributo, organizando-os em uma lista de ordem decrescente, desde o mais até o menos relevante para a classificação. Porém, visando estabelecer o subconjunto de atributos que será utilizado na classificação, faz-se necessário determinar um ponto de corte nessa lista, mantendo no subconjunto apenas os atributos posicionados acima dele.

Uma vez que os métodos de filtragem supramencionados não oferecem nenhuma informação sobre em que ponto deve ser feito o corte, uma busca exaustiva foi realizada para determinar o melhor ponto de corte possível em cada caso.

A busca é realizada adicionando-se atributos progressivamente no subconjunto. Inicialmente, o subconjunto é formado apenas pelo melhor atributo classificado pelo filtro. O desempenho desse subconjunto de atributos (no caso,

apenas um) é então avaliado por cada classificador, utilizando validação cruzada, conforme explicado detalhadamente no item 4.5. Em seguida, o segundo melhor atributo é incluído no subconjunto, que agora passava a contar com os dois melhores atributos, e esse novo subconjunto é avaliado por todos os classificadores novamente. Esse processo é repetido até que o subconjunto de atributos se iguale ao conjunto completo de atributos, finalizando assim a busca.

4.5 Validação cruzada

Todos os filtros e classificadores deste trabalho foram implementados na linguagem Java, em bibliotecas do Weka 3.6.6, que é uma coleção de algoritmos de aprendizado de máquina. Para avaliar o desempenho dos subconjuntos, utilizou-se a validação cruzada *10-fold*, que permitiu estimar as taxas de acerto dos classificadores gerados.

Nesse procedimento, as amostras são divididas aleatoriamente em 10 subconjuntos de aproximadamente o mesmo tamanho e mesma distribuição de classes. Um desses subconjunto é separado como conjunto de teste e os demais são usados como conjuntos de treino. Esse processo é então repetido 10 vezes, alternando-se o conjunto de teste usado em cada vez. Para produzir o resultado final das taxas de acerto, a média das 10 alterações é calculada.

Para minimizar a variação entre os testes, o processo de validação cruzada foi repetido quinze vezes, sempre dividindo os dados aleatoriamente. Ao final, a média dos resultados das repetições foi o indicador de taxa de acerto utilizado.

4.6 Considerações finais do capítulo

Neste capítulo descreveu-se como se deu a realização do experimento proposto para o trabalho. Desde a coleta das mudas em uma estufa comercial de Holambra, SP, passando por todo o processo de tratamento e processamento das imagens, a extração de atributos e a consecutiva filtragem e avaliação das taxas de acerto obtidas em cada caso. O capítulo a seguir tratará de expor e discutir os resultados obtidos.

5 RESULTADOS E DISCUSSÕES

Este capítulo apresenta e discute os principais resultados obtidos neste trabalho. Face à sua originalidade, o autor se sentiu estimulado a submetê-los a uma revista científica de importância na área agrícola, a *Computer and Electronics in Agriculture*² (COMPAG). Sendo assim, esse capítulo baseia-se no supramencionado artigo de Silva et al. (2013), que foi aceito e publicado pela revista em setembro de 2013.

5.1 Da filtragem

A Tabela 1 apresenta os resultados obtidos pelos métodos de filtragem adotados no trabalho, que foram aplicados ao conjunto de 26 atributos mencionados no item 4.3. Conforme discutido, os métodos Chi-quadrado, *Info Gain* e *Gain Ratio* ordenam os atributos em ordem decrescente de relevância para o aprendizado da tarefa de classificação. Já o CFS conta um critério de corte interno e por isso não ordena a lista completa de atributos como fazem os outros métodos.

A área da muda e a área do círculo ficaram em primeiro e segundo lugar, em termos de relevância, para todos os filtros considerados. Excetuando-se eles, nenhum outro atributo foi consensualmente classificado pelos quatro filtros, demonstrando que não há uma distinção clara quanto a relevância da maioria dos atributos extraídos. Este era um resultado previsto, em alguma extensão, pois tão pouco existe consenso entre os especialistas sobre a relevância dos atributos ao realizar a inspeção visual, conforme discutido nas motivações do trabalho.

Ainda assim, pode-se notar semelhanças marcantes nos resultados oferecidos pelos métodos Chi-quadrado e *Info Gain*, e algumas outras semelhanças menos evidentes, mas ainda assim presentes numa análise mais abrangente. A distância entre a base e o centroide e a sexta, a sétima, a décima e a décima primeira subáreas foram classificadas de forma semelhante por todos os métodos utilizados. Da mesma forma, a terceira, a quarta e a décima sexta subáreas foram todas classificadas nas três últimas posições da tabela.

² Classificação B1 para Engenharias IV e A2 para Ciências Agrárias segundo classificação do sistema WebQualis. Acessado em 10/2013.

Tabela 1 - Ordem de relevância dos atributos segundo cada filtro utilizado.

Posição	Chi-quadrado	Info Gain	Gain Ratio	CFS
1ª	Área muda	Área muda	Área muda	Área muda
2ª	Área círculo	Área círculo	Área círculo	Área círculo
3ª	Dist. base-centroide	Dist. base-centroide	subárea06	Ocupação círculo
4ª	Comprimento	Comprimento	Comprimento	Dist. base-centroide
5ª	Área retângulo	Área retângulo	Subárea 10	Subárea 06
6ª	Subárea 06	Subárea 06	Dist. base-centroide	Subárea 07
7ª	Subárea 10	Subárea 10	Área retângulo	Subárea 10
8ª	Largura	Largura	Subárea 11	Subárea 11
9ª	Subárea 07	Subárea 11	Elongamento	Subárea 15
10ª	Subárea 11	Subárea 07	Subárea 15	-
11ª	Subárea 01	Subárea 01	Largura	-
12ª	Elongamento	Subárea 12	Subárea 07	-
13ª	Subárea 12	Subárea 13	Subárea 13	-
14ª	Subárea 13	Elongamento	Subárea 01	-
15ª	Subárea 08	Subárea 08	Subárea 14	-
16ª	Subárea 05	Dist. centroide-círculo	Dist. centroide-círculo	-
17ª	Dist. centroide-círculo	Subárea 05	Subárea 02	-
18ª	Subárea 14	Subárea 14	Subárea 09	-
19ª	Ocupação círculo	Ocupação círculo	Subárea 12	-
20ª	Ocupação retângulo	Ocupação retângulo	Subárea 05	-
21ª	Subárea 09	Subárea 02	Subárea 08	-
22ª	Subárea 02	Subárea 09	Ocupação retângulo	-
23ª	Subárea 15	Subárea 15	Ocupação círculo	-
24ª	Subárea 04	Subárea 04	Subárea 04	-
25ª	Subárea 16	Subárea 16	Subárea 16	-
26ª	Subárea 03	Subárea 03	Subárea 03	-

Algumas discrepâncias também chamam a atenção. A ocupação do círculo, terceiro melhor atributo segundo a avaliação do CFS, não superou a 19ª posição em nenhum dos três métodos de *feature ranking* adotados. Já o comprimento, em contrapartida, não foi selecionado pelo CFS, apesar de ter sido unanimemente posicionado na 4ª colocação pelos outros filtros. A décima quinta subárea, em contrapartida, foi considerada praticamente irrelevante segundo o Chi-quadrado e o *Info Gain* (23ª posição em ambos os casos), mas se dividiu entre a décima e nona posições quando avaliada pelo *Gain Ratio* e o CFS, respectivamente.

Esses resultados demonstram que a relevância atribuída a cada atributo pode variar consideravelmente dependendo da métrica adotada. Apesar disso, são poucos os casos em que essa diferença de relevância ocorre de maneira acentuada. No geral, os atributos tendem a apresentar distribuições semelhantes, subindo ou descendo apenas algumas posições quando o método de filtragem é substituído.

5.2 Da avaliação dos subconjuntos

As Figuras Figura 14, Figura 15 e Figura 16 demonstram graficamente os resultados obtidos pelo método de definição de subconjuntos discutido no item 4.4. Observando-se as figuras, nota-se que o *Logistic Regression* foi o classificador que apresentou o melhor desempenho dentre todos os considerados. Pode-se ver nas figuras que as curvas de taxa de acerto geradas por esse classificador foram suaves, com algumas inflexões na sua extensão. O pico das suas taxas de acerto, considerando-se todos os métodos de filtragem que avaliou, foi atingido quando de 12 a 17 atributos estavam contidos no subconjunto e as taxas de acerto caíram gradativamente conforme mais atributos foram adicionados. A seleção de atributos não só diminuiu a quantidade de atributos necessários para a classificação para quase metade da inicial, mas também aumentou a taxa de acerto do *Logistic* em até 2,8 % quando comparada à obtida utilizando-se o conjunto completo de atributos.

Figura 14 – Gráfico da variação da taxa de acerto obtida ao se variar a quantidade de atributos incluídos no subconjunto. A ordem adotada segue o critério de relevância estabelecido pelo Chi-quadrado.

Fonte: Autor

O *Random Forest* mostrou-se o segundo melhor, com desempenho similar ao do MLP. As suas curvas de taxas de acerto tiveram um comportamento regular dentre todos os filtros considerados: elas começam na menor porcentagem de acerto e sobem exponencialmente até um patamar quase invariante. As taxas de

acerto praticamente não mostraram redução conforme o subconjunto foi acrescido de mais e mais atributos, mantendo-se constante até mesmo quando se atingiu o conjunto completo de atributos. Esse comportamento é consistente com a teoria por trás dessa técnica de aprendizado. É criado um grande número de árvores decisão e a resposta fornecida baseia-se na opção mais votada dentre diversas árvores. Dessa forma, o RF evita a hiperespecialização frente ao conjunto de instâncias fornecidas e, por conseguinte se torna menos sensível a atributos irrelevantes ou redundantes. Ainda assim, esse classificador também se beneficia da seleção de atributos, pois o esforço computacional adicional necessário para se extrair um grande número de atributos das imagens não é compensado por melhorias nas taxas de acerto.

Figura 15 - Gráfico da variação da taxa de acerto obtida ao se variar a quantidade de atributos inclusos no subconjunto. A ordem adotada neste caso segue o critério de relevância estabelecido pelo *Information Gain*.

Fonte: Autor

Como mencionado, o *Multilayer Perceptron* forneceu o terceiro melhor resultado. Suas curvas de acurácia se comportaram de forma muito análoga às obtidas com o LR, porém com desempenho inferior e quedas menos contínuas nas taxas de acerto conforme a segunda metade dos atributos foi sendo inserida no subconjunto. Uma explicação para esse comportamento poderia ser a complexidade intrínseca desse algoritmo (que cresce exponencialmente conforme cresce o número

de atributos na entrada), fazendo com que o resultado seja mais sensível à quantidade de atributos adicionados ao subconjunto.

Figura 16 - Gráfico da variação da taxa de acerto obtida ao se variar a quantidade de atributos incluídos no subconjunto. Este último gráfico da sequência segue a ordem estabelecida pelo *Gain Ratio*.

Fonte: Autor

Nos resultados obtidos pelo 5NN, nota-se que suas curvas de acurácia apresentaram diversas inflexões, e que seu desempenho caiu mais rapidamente quando comparado com as demais a partir do momento em que mais de 10 a 12 atributos foram incluídos no subconjunto. Esse comportamento oscilante indica a alta sensibilidade desse classificador para a ocorrência de atributos redundantes ou irrelevantes, como já era esperado. Os ganhos na taxa de acerto, obtidos mediante a seleção de atributos, chegaram a até 8,8%.

O C4.5, que cria uma árvore de decisão como modelo de classificação, foi o segundo pior método para a classificação das mudas. Curiosamente, o comportamento de suas curvas de acurácia foi consideravelmente distinto daquele observado pelas curvas do RF, que também se baseia em árvores de decisão para realizar a classificação e que, analogamente ao C4.5, também conta com uma abordagem preocupada em atribuir maior relevância para atributos que melhor discriminem entre as classes. Além disso, as taxas de acerto obtidas pelo C4.5 são até 7,6 % menores do que as do RF em alguns pontos, sendo 4,7 % inferiores na

média. Finalmente, o desempenho do C4.5 apresentou uma leve queda conforme mais atributos foram sendo adicionados ao subconjunto indicando, assim, uma menor resistência desse algoritmo a ocorrência de atributos não relevantes. Uma possível explicação para essas diferenças de comportamento entre o C4.5 e o RF é o fato do primeiro contar apenas com uma árvore para realizar a classificação, sendo assim menos provável que ele consiga extrair dos dados tanta informação quanto o segundo. Não obstante, a hiperespecialização também tende a ocorrer com maior facilidade.

Finalmente, as taxas de acerto do *Naïve Bayes* foram as piores na maioria dos casos. Suas curvas de acurácia se comportaram de maneira diferente, de acordo com o método de filtragem utilizado. No Chi-quadrado e no *Info Gain* o desempenho apresentou queda para os cinco primeiros atributos e em seguida subiu para os próximos cinco, oscilando a partir de então com leve tendência de queda. Já no *Gain Ratio*, o desempenho subiu para os cinco primeiros atributos e oscilou a partir do sexto, terminando numa queda total de 2,9% de acerto. O comportamento muito similar das taxas de acerto desse classificador para os dois primeiros filtros pode ser explicado pela ordem quase idêntica que atribuíram a relevância dos atributos. Em contrapartida, o filtro *Gain Ratio* apresentou maiores diferenças sobre a importância de cada atributo, justificando o comportamento distinto da curva de acurácia do classificador para esse caso.

A partir das análises realizadas nesse item, pode-se entender que o comportamento das curvas de acurácia, salvo algumas exceções, foi crescente a princípio e, conforme os atributos eram adicionados, eventualmente atingiram um valor máximo e prosseguiram em queda até não haver mais atributos de fora. Desse comportamento podemos entender que, no geral, a seleção de atributos foi capaz de melhorar as taxas de acerto observadas, porém a definição de quais e quantos atributos devem ser incluídos no subconjunto a ser utilizado não é trivial, variando de acordo com o filtro ou classificador escolhidos.

5.3 Da comparação entre os classificadores

A Figura 17 resume o resultado de cada filtro para cada classificador, considerando a melhor taxa de acerto obtida dentre todas as avaliadas nas diversas configurações de subconjunto. Além disso, os resultados do CFS também foram

incluídos nesse momento uma vez que, pelo CFS realizar a definição do subconjunto automaticamente, não se adequavam a análise exaustiva discutida no item 5.2.

Figura 17 – Taxas finais de acerto obtidas antes e após a seleção de atributos para todos os casos analisados. Os números acima das barras indicam a quantidade de atributos no subconjunto utilizado.

Fonte: Autor

Em termos de aumentos na acurácia, alguns classificadores se beneficiaram mais da seleção de atributos do que outros. O maior ganho que se obteve veio do 5NN, que antes da seleção era o pior classificador dentre todos os outros, e depois dela passou a ter um desempenho comparável ao dos melhores. O *Logistic Regression*, C4.5 e o *Naïve Bayes* apresentaram alguma melhoria, enquanto o RF permaneceu quase inalterado.

O CFS foi particularmente eficaz para o 5NN e também para o *Naïve Bayes*, o que é consistente com seu funcionamento uma vez que ambos os classificadores são altamente sensíveis à presença de atributos redundantes, e uma das principais preocupações do idealizador do CFS (HALL, 1999) é a de removê-los.

Em termos de dimensionalidade, a quantidade de atributos necessária para a classificação das mudas de violeta foi consideravelmente reduzida em todos os casos, e excetuando-se a um deles, veio sempre acompanhada de melhorias na taxa de acerto.

O LR apresentou o melhor desempenho para a classificação das mudas e, após a seleção de atributos, reduziu-se em 10 a quantidade de atributos que

necessita para realizar essa classificação ainda com um ganho de 2,8% na sua taxa de acerto.

A acurácia do RF não apresentou melhorias significativas após a seleção de atributos. Ainda assim, a quantidade de atributos necessária para a classificação foi reduzida para menos da metade da inicial e, em três casos de quatro, contou de qualquer maneira com uma pequena melhora.

5.4 **Considerações finais do capítulo**

A combinação de 4 técnicas de filtragem e 6 classificadores forneceu resultados que auxiliaram a análise da eficácia das técnicas de seleção de atributo para a classificação das mudas agrícolas. Apesar de não haver consenso entre os métodos sobre quais atributos são mais relevantes para a classificação, diversas semelhanças puderam ser destacadas. O comportamento dos classificadores conforme mais atributos eram considerados no conjunto de treinamento foi semelhante na maioria dos casos com o aumento da taxa de acerto inicialmente, seguida de decaimento até que todos os atributos fossem inclusos. A comparação entre as diversas configurações finais obtidas mostrou que há melhorias no desempenho do sistema quando a seleção de atributos é realizada. No capítulo a seguir serão extraídas as conclusões sobre o trabalho realizado e os resultados obtidos.

6 CONCLUSÕES

Esse trabalho visou avaliar a influência e eficácia de técnicas de seleção de atributos quando aplicadas a classificação visual de mudas de violeta. Por meio da avaliação e seleção dos atributos mais relevantes para o aprendizado da tarefa em questão, procurou-se assegurar que toda a informação necessária à classificação fosse capturada pelo sistema, descartando-se, assim, atributos irrelevantes ou redundantes para o aprendizado.

Inicialmente, um grande conjunto de atributos foi estabelecido a partir das imagens das mudas. Em seguida, diversas configurações de subconjuntos foram geradas, utilizando-se diversas técnicas de filtragem de atributos. Nos casos cabíveis, realizou-se uma busca exaustiva para definir qual configuração apresentava o melhor desempenho para a classificação. Cada subconjunto possível foi avaliado com a ajuda de alguns dos algoritmos de aprendizado supervisionado mais difundidos na literatura, estimando as taxas de acerto que produziam, através de repetidas operações de validação cruzada.

A comparação entre as taxas de acerto obtidas antes e após a seleção de atributos mostrou que a utilização dessas técnicas trouxe ganhos de acurácia de em todos os casos analisados, sendo que a combinação do RF com o CFS foi a única exceção observada, mas que, ainda assim, sofreu queda de apenas 0,58% e usou somente nove dos vinte e seis atributos inicialmente considerados. Dessa forma, pode-se concluir que o uso de técnicas de seleção de atributos mostrou-se vantajoso para a inspeção visual de mudas de violeta, auxiliando na incorporação do conhecimento tácito envolvido na classificação das mudas e gerando ganhos de acurácia de até 8,8 %. Além disso, a seleção de atributos reduziu de 26 para 11 a quantidade média de atributos usados na classificação, o que simplifica o processo de extração de atributos das imagens e facilita a compreensão dos classificadores gerados.

Os classificadores 5NN e *Naïve Bayes* foram os mais suscetíveis à presença de atributos irrelevantes ou redundantes, sendo os que mais se beneficiaram da seleção de atributos. O RF, por sua vez, mostrou-se o menos suscetível à presença desse tipo de atributos, mas também se beneficiou da seleção de atributos pela grande redução do número de atributos necessários para atingir resultados muito próximos dos originais.

Nos experimentos realizados, a melhor combinação entre subconjunto e classificador foi a do LR com o Chi-quadrado, utilizando-se os primeiros 16 atributos segundo a ordem da Tabela 1. Essa configuração atingiu uma taxa de acerto de 80,0%, sendo assim uma boa candidata para a automação da classificação de mudas de violeta.

Quadro 1 - Resumo das variações nas taxas de acerto após a seleção de atributos.

Classificador Téc. seleção Nº atributos Var. acerto	LR			
	Chi-Quadrado	Info Gain	Gain Ratio	CFS
	16	17	12	9
	2,78%	2,62%	1,96%	0,19%
Classificador Téc. seleção Nº atributos Var. acerto	RF			
	Chi-Quadrado	Info Gain	Gain Ratio	CFS
	10	13	13	9
	0,70%	0,51%	0,41%	-0,58%
Classificador Téc. seleção Nº atributos Var. acerto	MLP			
	Chi-Quadrado	Info Gain	Gain Ratio	CFS
	10	9	13	9
	3,38%	3,36%	3,09%	2,13%
Classificador Téc. seleção Nº atributos Var. acerto	5NN			
	Chi-Quadrado	Info Gain	Gain Ratio	CFS
	10	10	12	9
	7,37%	7,37%	5,99%	8,78%
Classificador Téc. seleção Nº atributos Var. acerto	C4.5			
	Chi-Quadrado	Info Gain	Gain Ratio	CFS
	6	5	3	9
	2,42%	2,42%	2,97%	1,29%
Classificador Téc. seleção Nº atributos Var. acerto	Naïve Bayes			
	Chi-Quadrado	Info Gain	Gain Ratio	CFS
	10	10	5	9
	1,14%	1,14%	2,91%	4,80%

6.1 Dificuldades encontradas

Dentre as principais dificuldades enfrentadas para o desenvolvimento deste trabalho, cabe destacar a de compreensão da dinâmica de uso das bibliotecas de funções do OpenCV. Apesar de dispor de um bom volume de documentação *online*, ela não abrange alguns aspectos importantes sobre a manipulação dos dados das imagens e muitas vezes apresenta funções e operadores pouco intuitivos e

praticamente não explicados, apenas mencionados. Isto é natural por se tratar de software livre, que evolui de forma descentralizada.

6.2 Contribuições

As contribuições que esse trabalho oferece podem ser vistas na listagem a seguir:

- O trabalho elenca uma lista de 26 atributos que podem ser utilizados para modelagem de mudas agrícolas, levantados após diversas interações com especialista do ramo. Esse levantamento facilita trabalhos futuros que desejem lidar com esse tipo de produto agrícola, permitindo maiores evoluções futuras nessa linha de pesquisa;
- Base de dados de imagens de mudas já rotuladas disponibilizada online para a comunidade científica. Essa base já está sendo usada por outros autores dentro do próprio laboratório, como por exemplo o artigo de Bieleveld et al. (2013);
- A incorporação das tecnologias propostas em uma máquina funcional, gerando um produto para a sociedade. Maiores detalhamentos sobre esse desdobramento do trabalho são discutidos em Silva et al. (2013);
- A comparação de diversas técnicas de seleção de atributos e seu desempenho com os principais classificadores da literatura, evidenciando suas vantagens e desvantagens em cada caso.

6.3 Limitações

O trabalho realizado apresenta algumas limitações importantes. A primeira que se pode mencionar é o fato de que, conforme explicado no Capítulo 4, a detecção da raiz da muda depende da orientação desta para o lado direito da imagem, obrigatoriamente. Isso faz com que a análise não seja invariante a rotação, uma limitação relevante no caso de aplicações comerciais.

Além disso, os métodos de filtragem do tipo *feature ranking* não oferecem critério de corte específico, o que fez necessário realizar uma busca exaustiva entre todos os pontos de corte possíveis. Essa abordagem não é a ideal, pois caso o número de atributos ou a quantidade de filtros que se deseja avaliar seja muito

grande, a análise pode se tornar computacionalmente inviável. O escopo do trabalho realizado permitiu que não houvesse maiores preocupações com essa questão, mas é uma limitação que deve ser mencionada.

6.4 Trabalhos futuros

Entre os principais trabalhos futuros em continuidade ao realizado e que podem ser sugeridos, tem-se:

Deteção da raiz invariante a rotação: como citado no item 6.3, a ausência da capacidade de detectar a posição da raiz da muda independentemente de sua rotação é uma limitação importante do trabalho. Estudos que procurem realizar essa detecção podem tornar mais vantajosa a automação da análise das mudas de em aplicações comerciais da tecnologia. A simetria em relação ao eixo longitudinal da muda poderia ser explorada para se atingir esse objetivo.

Determinação do ponto de corte para filtros de *feature ranking*: também de acordo com o que foi discutido no item 6.3, a análise exaustiva e caso a caso sobre qual o melhor ponto de corte para os métodos de filtragem que apenas ordenam os atributos de acordo com sua relevância é uma solução prática mas não genérica, pois conjuntos muito grandes de atributos podem inviabilizar essa abordagem. Trabalhos futuros podem se dedicar a estabelecer critérios de corte automáticos, efetivamente transformando a análise individual em uma análise de seleção de subconjunto.

Avaliação por mais de um especialista: este trabalho foi desenvolvido baseado na experiência de apenas um especialista, que foi escolhido pela sua comprovada competência, conforme comentado no item 4.1. Apesar da competência do especialista escolhido, ele realizou a categorização das mudas entre os quatro níveis de qualidade adotados de acordo com sua análise de cada uma delas. Uma vez que a fronteira entre essas classes é tênue, mesmo um especialista pode cometer equívocos. A análise das mudas por uma banca de especialistas, que atribuam notas a cada uma delas, pode gerar análises interessantes em trabalhos futuros, melhorando a informação base fornecida ao sistema para realizar a classificação e, possivelmente, aumentando as taxas de acerto obtidas.

6.5 Considerações finais

Técnicas de seleção de atributos se mostraram eficazes em fazer com que os classificadores produzam maiores taxas de acerto. Sua capacidade de filtrar atributos que agregam pouca ou nenhuma informação ao sistema foi nitidamente demonstrada pelos resultados obtidos no trabalho. Ainda assim, elas não eliminam a necessidade de se interagir com os especialistas no momento de definir quais atributos devem ser medidos para a classificação de produtos agrícolas. Essas técnicas apenas auxiliam o projetista do sistema a compreender melhor a relevância dos atributos obtidos, possivelmente permitindo que parte deles seja removida na versão final do sistema.

Finalmente, cabe ressaltar que apesar da pesquisa ter sido realizada no contexto de um trabalho de mestrado, o caráter inovadores dos resultados aqui obtidos permitiram sua publicação em periódico internacional, conforme mencionado no início do capítulo 5. O trabalho aqui desenvolvido também recebeu reconhecimento interno, sendo eleito o melhor trabalho de mestrado do II Workshop de Pós-Graduação da Área de Concentração “Engenharia da Computação”. Ademais disso, apesar de não abordada em detalhes no corpo deste documento, um dos desdobramentos deste trabalho foi a incorporação do sistema gerado numa máquina, capaz de realizar a classificação e separação das mudas automaticamente. Esse desdobramento vem alcançando uma repercussão importante, tanto do ponto de vista acadêmico, demonstrado por publicação dele decorrente, que foi escolhida como uma das oito melhores publicações do evento em que foi apresentada, como do ponto de vista de aplicação, originando diversas reportagens³ (OLIVETTE, 2013; REZENDE, 2013) e culminando com a aprovação, em setembro de 2013, do projeto PIPE 12/50974-9, que possibilitará continuar a transpor os resultados aqui discutidos em produtos reais, úteis à sociedade.

³ OLIVETTE, C. Integração de startups estimula a inovação. O Estado de São Paulo, São Paulo, 2 de jun. 2013. Oportunidades, p. 4.

REZENDE, R. Negócios já surgem na faculdade. O Estado de São Paulo, São Paulo, 28 de ago. 2013. Economia, p. B5A.

Universitários desenvolvem técnicas de cultivo de violeta. São Paulo: EPTV, 2012. Exibido na Rede Globo de televisão, Globo Rural, 31 de dez. 2012.

REFERÊNCIAS

- BASHEER, I. A.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. **Journal of microbiological methods**, [S.l.], v. 43, n. 1, p. 3-31, 2000.
- BLASCO, J. et al. Development of a machine for the automatic sorting of pomegranate (*Punica granatum*) arils based on computer vision. **Journal of Food Engineering**, [S.l.], v. 90, n. 1, p. 27-34, 2009.
- BIELEVELD et al. Symbolic regression via linear genetic programming for pot plant seedling classification. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 9., 2013, Mato Grosso. **Anais...** No prelo.
- BRADSKI, G.; KAEHLER, A. **Learning OpenCV: Computer vision with the OpenCV library**. O'reilly, 2008.
- BREIMAN, L. Random forests. **Machine learning**, [S.l.], v. 45, n. 1, p. 5-32, 2001.
- BRONS, A. et al. Plant grading by vision using neural networks and statistics. **Computers and electronics in agriculture**, [S.l.], v. 9, n. 1, p. 25-39, 1993.
- CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. In: International conference on Machine learning, 23., 2006. **Proceedings...** ACM, 2006. p. 161-168.
- COSTA, C. et al. Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. **Food and Bioprocess Technology**, [S.l.], v. 4, n. 5, p. 673-692, 2011.
- COVER, T.; HART, P. Nearest neighbor pattern classification. In: Information Theory. **IEEE Transactions on**, v. 13, n. 1, p. 21-27, 1967.
- CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of control, signals and systems**, [S.l.], v. 2, n. 4, p. 303-314, 1989.
- DASH, M. et al. Feature selection for clustering-a filter solution. In: IEEE International Conference on Data Mining, 3., 2002, Florida. **Proceedings...** 2003. p. 115-122.
- FAUST, J. E.; HEINS, R. D. Modeling inflorescence development of the African Violet (*Saintpaulia ionantha* Wendl.). **Journal of the American Society for Horticultural Science**, [S.l.], v.119, n. 4, p. 727-734, 1994.
- GOLNABI, H.; ASADPOUR, A. Design and application of industrial machine vision systems. **Robotics and Computer-Integrated Manufacturing**, [S.l.], v. 23, n. 6, p. 630-637, 2007.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **The Journal of Machine Learning Research**, [S.I.], v. 3, p. 1157-1182, 2003.

HALL, M. A. **Correlation-based feature selection for machine learning**. 1999. 178 p. Tese (Doutorado) - The University of Waikato, Nova Zelândia.

HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. **IEEE Transactions on Knowledge and Data Engineering**, [S.I.], v. 15, n. 6, p. 1437-1447, 2003.

HALLORAN, J. Classification: Naïve Bayes vs Logistic Regression. Disponível em: <<http://melodi.ee.washington.edu/~halloj3/classification.pdf>>. Acesso em: 26 jan 2014. 2009.

HASTIE, T.; TIBSHIRANI R.; FRIEDMAN, J. J. H. **The elements of statistical learning**. New York: Springer, 2001.

HUANG, Y. J.; LEE, F. F. An automatic machine vision-guided grasping system for Phalaenopsis tissue culture plantlets. **Computers and Electronics in Agriculture**, v. 70, n. 1, p. 42-51, 2010.

LEWIS, J. P. Fast template matching. In: **Vision Interface**. Canadian Image Processing and Pattern Recognition Society, 1995. p. 15-19.

JI, Q.; SINGH, S. Automated visual grading of vegetative cuttings. In: **Photonics East'96**. International Society for Optics and Photonics, 1996. p. 88-99.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant features and the subset selection problem. **ICML**, [S.I.], v. 94, p. 121-129, 1994.

JORDAN, A. On discriminative vs. generative classifiers: A comparison of logistic regression and Naïve bayes. **Advances in neural information processing systems**, [S.I.], v. 14, p. 841, 2002.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial intelligence**, [S.I.], v. 97, n. 1, p. 273-324, 1997.

Selection of relevant features in machine learning. Defense Technical Information Center, 1994.

MALAMAS, E. N. et al. A survey on industrial vision systems, applications and tools. **Image and vision computing**, v. 21, n. 2, p. 171-188, 2003.

MATTSSON, A. Predicting field performance using seedling quality assessment. **New Forests**, [S.I.], v. 13, n. 1-3, p. 227-252, 1997.

MITAL, A.; GOVINDARAJU, M.; SUBRAMANI, B. A comparison between manual and hybrid methods in parts inspection. **Integrated Manufacturing Systems**, v. 9, n. 6, p. 344-349, 1998.

MITCHELL, T. M. **Machine Learning**. Burr Ridge, IL: McGraw Hill, 1997. Vol. 45.

OLIVETTE, C. Integração de startups estimula a inovação. **O Estado de São Paulo**, São Paulo, 2 de jun. 2013. Oportunidades, p. 4.

PARSONS, N. R.; EDMONDSON, R. N.; SONG, Y. Image analysis and statistical modelling for measurement and quality assessment of ornamental horticulture crops in glasshouses. **Biosystems Engineering**, [S.l.], v. 104, n. 2, p. 161-168, 2009.

PARSONS, N. R.; EDMONDSON, R. N.; GILMOUR, S. G. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, [S.l.], v. 55, n. 4, p. 507-524, 2006.

QUINLAN, J. R. **C4. 5: programs for machine learning**. San Mateo: Morgan Kaufmann, 1993. Vol. 1.

REZENDE, R. Negócios já surgem na faculdade. **O Estado de São Paulo**, São Paulo, 28 de ago. 2013. Economia, p. B5A.

RISH, I. An empirical study of the Naïve Bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. 2001. p. 41-46.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533-536, 1986.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, [S.l.], v. 23, n. 19, p. 2507-2517, 2007.

SILVA, L. O. L. A. et al. Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings. **Computers and Electronics in Agriculture**, v. 97, p. 47-55, 2013.

SILVA, L. O. L. A. et al. Automated visual quality sorting of agricultural seedlings. In: Workshop de Visão Computacional, 9., 2013, Rio de Janeiro. **Anais...** Rio de Janeiro: FGV, 2013. Disponível em: <http://iris.sel.eesc.usp.br/wvc/Anais_WVC2013/Oral/3/3.pdf>. Acesso em: 01 nov. 2013.

SINGH, K.; GUPTA, I.; GUPTA, S. Classification of bamboo plant based on digital image processing by Central moment. In: **Image Information Processing (ICIIP), 2011 International Conference on**. IEEE, 2011. p. 1-5.

TIMMERMANS, A. J. M.; HULZEBOSCH, A. A. Computer vision system for on-line sorting of pot plants using an artificial neural network classifier. **Computers and electronics in agriculture**, [S.l.], v. 15, n. 1, p. 41-55, 1996.

TRIER, O. D.; JAIN, A. K. Goal-directed evaluation of binarization methods. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 17, n. 12, p. 1191-1201, 1995.

UNAY, D et al. Automatic grading of Bi-colored apples by multispectral machine vision. **Computers and Electronics in Agriculture**, v. 75, n. 1, p. 204-212, 2011.

WANG, Z.; ZHANG, D. Progressive switching median filter for the removal of impulse noise from highly corrupted images. **Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on**, v. 46, n. 1, p. 78-80, 1999.

WITTEN, I. H. et al. **Weka: Practical machine learning tools and techniques with Java implementations**. 1999.

YU, L.; LIU, H. Efficient feature selection via analysis of relevance and redundancy. **The Journal of Machine Learning Research**, [S.I.], v. 5, p. 1205-1224, 2004.

ZHENG, C.; SUN, D.; ZHENG, L. Recent developments and applications of image features for food quality evaluation and inspection—a review. **Trends in Food Science & Technology**, [S.I.], v. 17, n. 12, p. 642-655, 2006.