Aprendizado Ativo para Classificação do Vigor de Sementes de Soja

Douglas Felipe Pereira*, Pedro Henrique Bugatti*, Priscila Tiemi Maeda Saito*†
*Departamento de Computação, Universidade Tecnológica Federal do Paraná (UTFPR-CP)
†Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
Email: douglaspereira@alunos.utfpr.edu.br, {pbugatti, psaito}@utfpr.edu.br

Abstract—**The task of providing a high quality grain (e.g. soybean) to the farmer is a key challenge of the agrobusiness field. To achieve such quality considering soybean seeds it is applied the so-called tetrazolium test. This test provides an acurate diagnosis of the damages found in the seed, such as lacerations caused by insects, mechanical damages or high rates of humidity. These damages cause a considerable quality reduction and directly impact in the seed vigor. Some traditional machine learning methods were applied to the context of seed crops, in order to automatic classify the seed vigor. However, the great majority of the researches use the traditional supervised learning paradigm. Thus, in this paper we proposed to exploit the active learning paradigm to perform the classification of the seed vigor, derived from the tetrazolium test.**

***Keywords*-active learning, image analysis, image processing, classification, soybean seeds**

Resumo—**Oferecer grãos de qualidade ao produtor é um dos desafios do setor agroindustrial de grãos como os da soja. Para atingir tal qualidade em sementes de soja aplica-se amplamente o chamado teste de tetrazólio. Tal teste visa definir o vigor da semente, bem como apontar os tipos de danos encontrados na semente, como por exemplo os causados por umidade, por percevejo ou mecânicos. Algumas metodologias no contexto de aprendizado de máquina tradicional foram aplicadas a culturas de sementes, visando a classificação automática do vigor das mesmas. No entanto, a grande maioria das propostas utiliza o aprendizado supervisionado tradicional. Dessa forma, o presente trabalho visa explorar uma metodologia de aprendizado ativo para a classificação do vigor de sementes de soja, oriundas do teste de tetrazólio.**

Palavras-chaves-**aprendizado ativo; análise de imagens; processamento de imagens; classificação; sementes de soja**

I. Introdução

A soja é uma das culturas mais importantes do setor agroindustrial mundial [1] e sua produção está entre as atividades econômicas que apresentaram crescimentos mais expressivos nas últimas décadas [2], devido a fatores como o sólido mercado internacional, a crescente demanda dos setores que a utilizam como matéria prima e da geração e oferta de tecnologias que viabilizam a expansão do produto.

Tecnologias aplicadas ao cultivo de grãos têm sido amplamente desenvolvidas atualmente [3]. Para o cultivo da soja é muito importante determinar o vigor das sementes que serão utilizadas para o plantio, pois é o vigor que determina o potencial fisiológico da semente em desenvolver uma boa plântula no campo [4], [5]. Diversos testes podem ser utilizados para a determinação do vigor das sementes de soja, dentre eles: envelhecimento acelerado, condutividade elétrica, crescimento de plântulas e o teste de tetrazólio [6].

Dentre os diversos tipos de testes para controle da qualidade das sementes adotados pelas indústrias do setor de replicação de sementes, o teste de tetrazólio se destaca pela rapidez, precisão e pelo número de informações geradas à respeito do lote de sementes. Além disso fornece um diagnóstico mais preciso das principais causas de redução da qualidade, como danos mecânicos, por percevejo ou por umidade, que são as causas mais comuns que podem afetar a semente [4]. Esse teste é realizado por meio de uma análise visual da semente por especialistas e demanda do especialista conhecimento intrínseco sobre os problemas que podem acontecer em uma semente. O especialista analisa milhares de sementes por dia, tornando o teste de tetrazólio um processo cansativo e altamente sucetível a erros.

Um sistema de captura e armazenamento de imagens poderia ser aplicado para o registro e anotação das sementes de soja analisadas pelo teste de tetrazólio. No entanto, esse processo pode gerar um conjunto muito grande de dados, inviabilizando a anotação manual de todas as imagens. Para esse problema, um método de aprendizado ativo pode ser aplicado.

Alguns métodos de aprendizado ativo [7], [8] podem selecionar um conjunto razoavelmente pequeno de amostras que contém amostras relevantes para a criação de um modelo de classificação e utilizar o especialista para verificação/correção de amostras classificadas incorretamente em poucas iterações de aprendizado. O resultado é um classificador que é capaz de realizar a correta rotulação das amostras restantes do conjunto.

Durante o aprendizado ativo, o classificador participa ativamente do seu próprio aprendizado, sugerindo rótulos para o especialista realizar uma correção em cada iteração. Um dos desafios do aprendizado ativo é descobrir quais amostras selecionadas serão mais relevantes para um rápido aprendizado, com maior acurácia e menos iterações e tempo computacional.

Diversas técnicas têm sido desenvolvidas e aplicadas no contexto de classificação de grãos [9]–[12]. No entanto, tais técnicas utilizam metodologias de aprendizado supervisionado tradicional. Nenhuma delas têm explorado a utilização de técnicas de aprendizado ativo.

Contribuições: Este trabalho apresenta uma abordagem de aprendizado mais efetiva e eficiente para classificação do vigor de sementes de soja, com base em um paradigma de aprendizado ativo, o qual considera uma pré-organização do

conjunto de dados, e posterior seleção das amostras mais informativas para o processo de aprendizado do classificador.

II. Metodologia

A metodologia proposta explora uma abordagem de aprendizado ativo, *Root Distance-Based Sampling* (RDS) [13], que realiza um pré-processamento, organizando os dados previamente. Posteriormente, a partir da pré-organização, a estratégia de seleção consiste em priorizar diversidade (amostras de todas as classes) e incerteza (amostras mais difíceis para o aprendizado do classificador). O processo de aprendizado torna-se mais rápido, uma vez que não requer a classificação e re-organização de todas as amostras do conjunto de dados.

A. Estratégia de Organização

A estratégia de organização consiste em realizar o agrupamento das amostras do conjunto de aprendizado e a organização das amostras de cada cluster com base na distância relativa à raiz do cluster correspondente. Considerando o agrupamento de um conjunto de dados em nc clusters $\mathcal{C}_i$, sendo nc o número de classes e $i = 1, ...nc$, o processo de organização constrói uma lista ordenada de amostras por cluster. São criadas nc listas $\mathcal{L}_i$, uma para cada cluster i. As amostras presentes em cada lista $\mathcal{L}_i$ são organizadas em ordem crescente de distância em relação à raiz r_i do cluster correspondente $\mathcal{C}_i$.

B. Estratégia de Seleção

Na primeira iteração, o especialista anota os rótulos das raízes dos clusters, as quais são utilizadas para o treinamento da primeira instância do classificador. Nas iterações subsequentes a instância do classificador atual seleciona um dado número de amostras de cada lista ordenada, de forma a abranger amostras de todas as classes (diversidade) mais rapidamente.

Seguindo a ordem das amostras em cada lista, a estratégia começa a explorar a ideia de incerteza, ao priorizar amostras mais próximas às raízes e selecionar apenas a amostra cujo rótulo difere do rótulo da raiz correspondente, de acordo com a classificação da instância atual do classificador. Caso os rótulos não sejam distintos, as próximas amostras da lista são analisadas. Se a condição não é encontrada nas demais amostras da lista, a amostra que apresenta a maior distância (últimas amostras) nesta lista é selecionada, uma vez que amostras mais distantes às raízes podem indicar amostras de fronteira entre clusters.

As amostras selecionadas, a cada iteração do aprendizado, são submetidas à verificação de um especialista que confirma e/ou corrige os rótulos fornecidos pela instância atual do classificador. Após a verificação, tais amostras anotadas são adicionadas ao conjunto de treinamento da iteração anterior e uma nova instância do classificador é criada. O processo de aprendizado continua até que o especialista esteja satisfeito com as acurácias.

III. Experimentos

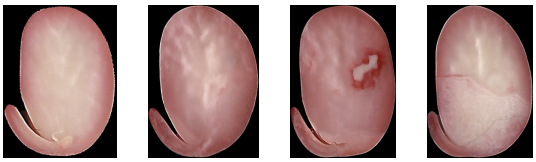
Nesta seção são descritos o conjunto de dados utilizado, bem como os cenários referentes aos experimentos realizados.

A. Conjunto de Sementes

O conjunto de dados utilizado consiste em 576 imagens de sementes de soja distribuídas de acordo com a Tabela I [11]. As classes estão nomeadas em uma tupla, onde o primeiro caractere refere-se ao nível de dano ocorrido na semente, os quais podem ser de 2 a 4. O segundo caractere ao tipo de dano (h-umidade, b-laceração por percevejo, m-mecânico), e o último caractere refere-se à porção da semente (E-externa ou I-interna). A classe perfeita, ou seja, sementes que não apresentam danos, é representada por pI (perfeita interna) e pE (perfeita externa). A Figura 1 ilustra os tipos de danos encontrados nas amostras do conjunto.

Tabela I
Classes e quantidade de amostras em cada classe

Classes	Amostras
2bE	20
2hI	68
2hE	51
3mI	17
3bI	54
3bE	50
3hI	38
3hE	50
4bI	74
4hE	15
pI	69
pE	70



(a) (b) (c) (d)

Figura 1. Tipos de danos em sementes de soja: (a) semente perfeita sem dano. (b) dano por umidade. (c) laceração por percevejo. (d) dano mecânico.

A descrição das imagens de soja foi realizada utilizando o descritor de cor *Border/Interior Classification* (BIC). Descritores baseados em cor obtiveram bons resultados na classificação das sementes de soja pelo teste de tetrazólio, devido à característica intrínseca do teste que difere a cor das sementes baseado nos danos ocorridos [11].

Para a classificação, diversos classificadores foram utilizados: *Optimum-Path Forest* (OPF), *Support Vector Machine* (SVM), *Random Forest* (RF), *MultiLayer Perceptron* (MLP), *NaiveBayes* (NB), *J48 Decision Trees* (J48) e *k-Nearest Neighboor* (IBK).

B. Cenários

A estratégia de organização utilizada foi baseada no agrupamento realizado pelo método Kmeans. A distância Euclidiana foi utilizada para ordenação das listas de amostras. Para efeito de comparação do método RDS abordado neste trabalho, foi utilizado o método de seleção aleatório que selecionou a mesma quantidade de amostras a cada iteração de maneira

aleatória e sem reposição no conjunto de treinamento. Os resultados discutidos na seção IV foram obtidos a partir da média dos experimentos executados 10 vezes com conjuntos de treinamento e teste gerados randomicamente e de modo balanceado a partir do conjunto de dados original. Para todas as execuções foram utilizados 80% das amostras disponíveis para treinamento do classificador e 20% para teste. Para a comparação entre os métodos foram utilizadas a medida da acurácia dos classificadores por iteração, a quantidade de amostras anotadas por iteração e a quantidade de classes conhecidas por iteração. Todos os classificadores utilizaram o mesmo agrupamento de amostras obtido pelo algoritmo de clusterização k-means.

IV. Resultados

Nesta seção são descritos os resultados dos experimentos realizados.

A Figura 2 mostra os valores de acurácia médias por iteração da execução do aprendizado ativo com os classificadores citados na seção anterior utilizando o método de organização e seleção do RDS. Como observado na Figura 2, o classificador OPF conseguiu alcançar acurácias maiores com menos iterações de aprendizado. Os classificadores SVM, RF e MLP apresentaram um desempenho similar nas primeiras iterações e o IBK obteve os piores resultados para a classificação nas primeiras iterações de aprendizado.

Para o aprendizado ativo, analisar as primeiras iterações é fundamental, pois o princípio é utilizar uma quantidade reduzida de amostras para maximizar a acurácia do classificador e reduzir o tempo computacional para treinamento do classificador.

O gráfico da Figura 3 ilustra que todos os classificadores obtiveram o mesmo comportamento com relação à quantidade de amostras corrigidas pelo especialista a cada iteração do aprendizado. Ao longo das iterações houve uma redução no número de amostras corrigidas pelo especialista, mostrando o avanço no aprendizado dos classificadores a cada iteração.

As Tabelas II-IV mostram a comparação dos classificadores que obtiveram os melhores resultados com o método RDS comparado com o método de seleção aleatório nas primeiras iterações de aprendizado. A quantidade de amostras selecionadas pelo RDS e pelo aleatório foi considerada a mesma por iteração para efeitos de comparação. Os classificadores selecionados para essa análise foram: OPF, SVM e RF.

Na Tabela II pode-se observar que a acurácia média inicial do classificador OPF com o método RDS foi maior do que a obtida pelo método de seleção aleatório e os outros métodos citados. Já analisando a Tabela III pode-se notar que o RDS apresenta uma taxa maior de anotação nas primeiras iterações, fato desejado, pois permite um melhor aprendizado ao classificador, além de diminuir consideravelmente tais anotações nas iterações posteriores.

Na Tabela IV é possível observar que nas iterações iniciais de aprendizado, o classificador OPF também consegue conhecer mais classes em menos iterações, isso faz com que a acurácia seja maior, pois a diversidade de classes é maior.

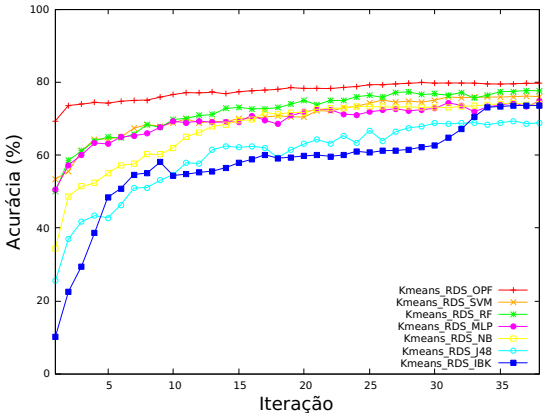


Figura 2. Acurácia por iteração para cada um dos classificadores.

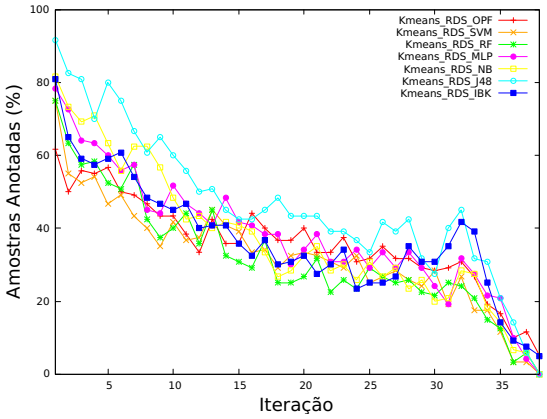


Figura 3. Amostras anotadas por iteração para cada um dos classificadores.

Outro fator a ser considerado é o tempo computacional para execução dos métodos, a partir do mesmo é possível verificar que o classificador OPF se destaca na classificação das amostras. Para os experimentos realizados com o método RDS foram mensurados o tempo demandado no agrupamento e na pré-organização das listas (i.e. tempo de pré-processamento), bem como o tempo do ciclo de aprendizado que corresponde à soma dos tempos de seleção, treinamento e teste por iteração de aprendizado. Os tempos obtidos estão na escala de segundos.

Para todos os classificadores o tempo de pré-processamento foi de $0,042s$, pois foi utilizado o mesmo agrupamento de amostras. Já o tempo do ciclo de aprendizado para o classificador OPF foi de $8,36s$, para o classificador SVM $9,87s$ e para o RF $109,96s$.

V. Conclusão

Neste trabalho foi apresentada uma abordagem de aprendizado ativo aplicado ao problema de classificação do vigor de sementes de soja pelo teste de tetrazólio.

A estratégia adotada pré-organiza as amostras do conjunto de dados, realizando o agrupamento das amostras e ordenando as amostras de cada cluster com base na distância relativa à raiz do cluster correspondente. A partir da pré-organização, a seleção consiste em priorizar amostras que apresentem

Tabela II
ACURÁCIAS MÉDIAS

Iter.	Kmeans_RDS_OPF	Rand_OPF	Kmeans_RDS_RF	Rand_RF	Kmeans_RDS_SVM	Rand_SVM
1	69,38 ± 1,13	66,48 ± 4,66	50,08 ± 4,96	44,23 ± 5,71	53,39 ± 4,53	41,86 ± 6,46
2	73,59 ± 1,17	71,83 ± 2,68	58,55 ± 3,82	55,42 ± 7,25	55,50 ± 4,84	54,57 ± 5,48
3	73,96 ± 1,87	73,21 ± 3,16	61,10 ± 4,17	61,01 ± 5,70	61,10 ± 3,18	59,83 ± 6,02
4	74,42 ± 2,50	74,36 ± 2,42	63,89 ± 4,21	63,55 ± 3,68	64,40 ± 2,46	62,03 ± 4,18
5	74,23 ± 1,89	74,40 ± 2,48	64,91 ± 4,03	65,84 ± 3,93	64,40 ± 4,01	63,13 ± 4,34

Tabela III
AMOSTRAS ANOTADAS (%)

Iter.	Kmeans_RDS_OPF	Rand_OPF	Kmeans_RDS_RF	Rand_RF	Kmeans_RDS_SVM	Rand_SVM
1	61,66 ± 0,16	55,84 ± 0,21	75,00 ± 12,42	57,50 ± 17,32	75,01 ± 0,13	59,17 ± 0,18
2	49,99 ± 0,11	39,17 ± 0,05	63,33 ± 10,54	52,50 ± 12,45	54,99 ± 0,11	46,67 ± 0,16
3	55,85 ± 0,12	41,67 ± 0,09	57,50 ± 14,40	36,66 ± 15,81	52,51 ± 0,12	49,16 ± 0,13
4	55,00 ± 0,15	44,16 ± 0,11	58,33 ± 14,69	37,50 ± 9,00	54,17 ± 0,13	34,99 ± 0,05
5	56,67 ± 0,11	41,68 ± 0,17	52,50 ± 7,90	31,66 ± 10,24	46,66 ± 0,15	34,99 ± 0,10

Tabela IV
CLASSES CONHECIDAS

Iter.	Kmeans_RDS_OPF	Rand_OPF	Kmeans_RDS_RF	Rand_RF	Kmeans_RDS_SVM	Rand_SVM
1	8,00 ± 0,90	7,40 ± 1,30	8,00 ± 0,90	7,10 ± 1,00	8,00 ± 0,90	7,50 ± 1,20
2	10,60 ± 0,70	10,00 ± 1,20	10,70 ± 1,10	9,40 ± 0,70	10,50 ± 0,70	10,00 ± 1,00
3	11,50 ± 0,50	11,00 ± 0,70	11,20 ± 0,80	10,50 ± 1,10	11,10 ± 0,60	10,70 ± 0,80
4	11,70 ± 0,50	11,60 ± 0,50	11,30 ± 0,80	10,90 ± 1,10	11,60 ± 0,70	11,40 ± 0,50
5	11,80 ± 0,40	11,60 ± 0,50	11,60 ± 0,50	11,20 ± 0,90	11,60 ± 0,70	11,60 ± 0,50

maior diversidade (amostras de todas as classes) e incerteza (amostras mais difíceis de serem classificadas).

Os experimentos foram realizados utilizando diversos classificadores propostos pela literatura, a fim de verificar o desempenho de cada um deles, bem como o mais indicado para obtenção de maiores valores de acurácias mais rapidamente, menores interações com o especialista e menor tempo computacional.

A técnica de aprendizado ativo foi comparada e apresentou resultados superiores em relação ao método de seleção aleatória de amostras do conjunto de dados completo, utilizando cada um dos classificadores. Com relação à utilização dos classificadores, a técnica de aprendizado ativo utilizando o classificador OPF foi a que atingiu médias de acurácias maiores nas primeiras iterações. Esse resultado mostra que o classificador consegue aprender mais rapidamente sobre o modelo de classificação do conjunto de dados em menos iterações de aprendizado, fazendo com que o tempo computacional de aprendizado seja menor.

Para trabalhos futuros, pretende-se desenvolver novas técnicas de aprendizado ativo, bem como explorar diferentes técnicas de agrupamento de dados. Além disso, pretende-se avaliar a integração de técnicas de aprendizado ativo com técnicas de aprendizado semi-supervisionado.

AGRADECIMENTOS

Os autores gostariam de agradecer à CAPES, CNPq, Fundação Araucária, UTFPR e SETI pelo apoio financeiro.

REFERÊNCIAS

[1] M. C. M. Freitas, "A cultura da soja no brasil: O crescimento da produção brasileira e o surgimento de uma nova fronteira agrícola," *Enciclopédia Biosfera*, vol. 7, no. 12, 2011.

[2] M. H. Hirakuri and J. J. Lazzarotto, "O agronegócio da soja nos contextos mundial e brasileiro," 2014. [Online]. Available: http://www.infoteca.cnptia.embrapa.br/bitstream/doc/990000/1/Oagronegociodasojanoscontextosmundialebrasileiro.pdf

[3] M. G. F. Santanna, P. T. M. Saito, and P. H. Bugatti, "Content-based image retrieval towards the automatic characterization of soybean seed vigor," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, ser. SAC '14. New York, NY, USA: ACM, 2014, pp. 964–969. [Online]. Available: http://doi.acm.org/10.1145/2554850.2555007

[4] J. F. Neto, F. C. Krzyzanowski, and N. P. da Costa, "O teste de tetrazólio em sementes de soja," *Embrapa-CNPSo*, 1998.

[5] A. Hoffmaster, K. Fujimura, M. McDonald, and M. Bennett, "An automated system for vigor testing three-day-old soybean seedings." *Seed Science and Technology*, vol. 31, no. 3, 2003.

[6] J. M. Filho, A. L. P. Kikuti, and L. B. Lima, "Métodos para avaliação do vigor de sementes de soja, incluindo a análise computadorizada de imagens," *Revista Brasileira de Sementes*, vol. 31, no. 1, 2009.

[7] P. T. Saito, P. J. de Rezende, A. X. Falcão, C. T. Suzuki, and J. F. Gomes, "An active learning paradigm based on a priori data reduction and organization," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6086 – 6097, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414002012

[8] E. Lughofer, "Hybrid active learning for reducing the annotation effort of operators in classification systems," *Pattern Recognition*, vol. 45, no. 2, pp. 884–896, 2012.

[9] K. Kiratiratanapruk and W. Sinthupinyo, "Color and texture for corn seed classification by machine vision," in *Intelligent Signal Processing and Communications Systems (ISPACS), 2011 International Symposium on*, Dec 2011, pp. 1–5.

[10] L. A. I. Pabamalie and H. L. Premaratne, "A grain quality classification system," in *Information Society (i-Society), 2010 International Conference on*, June 2010, pp. 56–61.

[11] D. F. Pereira, P. T. M. Saito, and P. H. Bugatti, "An image analysis framework for effective classification of seed damages," in *31st ACM Symposium on Applied Computing (SAC)*, 2016, pp. 61–66.

[12] C. Xiao, C. Tao, X. Yi, W. Li, and T. Yuzhi, "Grain classification using hierarchical clustering and self-adaptive neural network," in *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, June 2008, pp. 4415–4418.

[13] P. T. Saito, C. T. Suzuki, J. F. Gomes, P. J. de Rezende, and A. X. Falcão, "Robust active learning for the diagnosis of parasites," *Pattern Recognition*, vol. 48, no. 11, pp. 3572 – 3583, 2015.