Identificação de Viabilidade de Leveduras Com Corante Vital Utilizando Histogramas de Palavras Visuais em Imagens Coloridas

Junior Silva Souza[1]
Hemerson Pistori[2]
Marney Pascoli Cereda[3]
Wesley Nunes Gonçalves [4]
Valguima Victoria Viana Aguiar Odakura[5]

Data de submissão: 29.12.2015
Data de aceitação: 17.08.2015

Resumo: Neste artigo é apresentada uma proposta de automatizar a classificação da viabilidade de leveduras da espécie *Saccharomyces cerevisae*, responsável pela produção comercial do etanol, usando como atributo a cor absorvida pelo corante vital azul de metileno. A metodologia é usada amplamente em usinas no Brasil e consiste em contar as células incolores que são consideradas viáveis, separando-as das coloridas de azul, consideradas não viáveis. O número de células viáveis por litro interfere no rendimento industrial. Como essa contagem é cansativa e resulta em erros, apresentamos como alternativa a técnica de visão computacional definida como o algoritmo *Bag-of-Word* (histograma de palavras visuais), bem como algumas extensões que agregam informações de cor e que podem ser adicionados ao algoritmo, isto porque o *Bag-of-Word* é usado para imagens em tons de cinza. Os atributos extraídos deste algoritmo com suas extensões foram utilizados para teste e treinamento de classificadores extraídos de algoritmos de aprendizagem supervisionada. Entre os algoritmos de classificação que usamos podemos

[1] paranadiodo@yahoo.com.br
[2]pistori@ucdb.br
[3]cereda@ucdb.br
[4]wesley.gonçalves@ufms.br
[5]Valguimadakura@ufgd.edu.br

destacar J48, SMO, Naives Bayes e IBk que estão implementados no ambiente WEKA. Os resultados foram analisados através do ANOVA que apresentou valor-p < 2e-16 indicando uma diferença estatística das técnicas analisadas. A técnica *Opponent* Color apresentou melhores resultados, representando um potencial de aplicação em condições reais das usinas.

Abstract: This article presents a proposal to automate the classification of viability for yeast *Saccharomyces cerevisiae* species, responsible for the commercial production of ethanol, using as attribute the color absorbed by vital dye methylene blue. The methodology is used widely in the Brazilians distilleries and consist in count the colorless cells that are considered feasible, separate them from colored blues considered non-viable. The number of viable cells per liter interferes with the industrial yields. Since this count is tiring and results in errors, an alternative technique is introduced as computer vision defined as the bag-of-word algorithm (histogram of visual words), and extensions that add color information's and that can be added to the algorithm, because the Bag-of-word is used for images in grayscale. The attributes extracted with this algorithm with its extensions were used for testing and training of classifiers extracted from supervised learning algorithm. Among the used algorithm we can highlight J48, SMO, Naives Bayes and IBK that are implemented in the WEKA environment. The results were analyzed using ANOVA that showed value-p < 2e-16 indicating a statistical difference of the techniques. The Opponent Color technique showed better results, representing a potential application in real of the distilleries.

.

1 Introdução

Com a implantação do Programa Nacional do Álcool (Proálcool) a cultura da cana-de-açúcar se tornou matéria prima atrativa no âmbito nacional. Em um panorama estadual a cultura da cana-de-açúcar foi inserida no Mato Grosso do Sul na década de 1980. Com o objetivo de introduzir uma nova fonte de combustível, grandes áreas cultiváveis e que antes eram utilizadas pela pecuária, passaram a ser utilizadas para o cultivo da cana. Devido ao aumento da área plantada e a elevação do número de indústrias de beneficiamento da cana, houve um reflexo na busca por tecnologias que melhorassem a qualidade e a produtividade nos processos de produção [13]. Entre os produtos extraídos da cana pela indústria destaca-se o etanol.

A técnica usada até hoje nas usinas brasileiras baseia-se na reciclagem das leveduras comerciais durante toda a safra de um ano agrícola. Para isso depois da fermentação terminada, as leveduras em suspensão são centrifugadas e tratadas com acido sulfúrico por 2 horas, lavadas com água limpa e novamente usadas na próxima fermentação [21].

A produção do etanol ocorre pela fermentação do caldo da cana diluído com água, denominado tecnicamente de mosto. Neste processo as leveduras da espécie *Saccharomyces cerevisae* são adicionadas ao mosto, consomem seu açúcar mosto e produzem etanol e gás-carbono [15]. A qualidade da produção do etanol está diretamente relacionada às atividades das leveduras [8],[10].

A viabilidade celular é o aspecto mais importante no rendimento da fermentação alcoólica, pois quanto maior o número de leveduras viáveis, melhor será o desempenho do processo. Como o ambiente das dornas causa estresse nas leveduras, é necessário controlar esse número durante todo o processo [9].

O método de microbiologia clássica que é usado para estabelecer viabilidade de leveduras é a contagem em placas de Petri pelo número de colônias. Por este método

volumes de amostras são homogeneizadas, diluídas serialmente em um diluente apropriado, plaqueadas sobre ou dentro de um meio sólido adequado, o qual é incubado sob temperatura apropriada por um determinado tempo, sendo então todas as colônias visíveis contadas manualmente ou por um contador de colônias eletrônico. Cada colônia crescida é proveniente de uma célula viável. A contagem é preferencialmente feita dentro do intervalo de 30 a 300 colônias para uma placa de 9 cm de diâmetro. O fator de diluição e o volume do inóculo são considerados para o cálculo do resultado final [20].

Portanto, para garantir à produção, é necessário efetuar análises em laboratório, considerando que as leveduras viáveis são responsáveis pela fermentação e as leveduras inviáveis não desempenham a fermentação como deveria [18]. Por sua importância a análise de viabilidade das leveduras é realizada rotineiramente nas usinas, mas necessitam de um método rápido o bastante para acompanhar o processo que dura no máximo 10 horas. A microbiologia clássica não atende esse quesito.

Nas atividades relacionadas às análises das leveduras, amostras são retiradas dos tanques contendo o mosto e examinadas em laboratório por um técnico responsável. Esta atividade consiste na identificação e contagem visual das leveduras, com o auxílio de um microscópio. Para a identificação das leveduras viáveis que são as que produzem álcool, as amostras são misturadas em água e corante azul de metileno (corante vital), que colore em azul as leveduras inviáveis [18]. Esta tarefa é suscetível a erros, pois este processo pode ser cansativo e subjetivo por ser uma atividade repetitiva e visual [12].

Atividades como a identificação e contagem das leveduras configuram-se em tarefas repetitivas e que podem ser realizadas automaticamente por meio de computadores. Como estas análises são feitas por meio de imagens, é possível automatizá-las por intermédio da visão computacional e aprendizagem supervisionada. Com técnicas de visão computacional é possível extrair atributos de imagens relacionados à cor, forma, textura e etc. Estas informações podem ser utilizadas para realizar a identificação e/ou o reconhecimento de

imagens, por meio de classificadores obtidos pelos algoritmos de aprendizagem supervisionada.

O objetivo deste trabalho foi avaliar o desempenho obtido pelo BoW(*Bag-of-Word)* [1] através da adição de informação de cor, e desenvolver uma aplicação para a identificação e contagem de leveduras separadamente em viáveis e inviáveis por sua cor azul ou incolor. No que se refere à informação de cor foram utilizadas 4 técnicas: CCV (*Color Coherence Vectors*) [2], CM (*Color Moments*) [3], BoC (*Bag-of-Color*) [4] e OpC (*Opponent Color)*[5]. O CCV extrai informação de cor através de regiões ou aglomerados de uma mesma cor. O CM extrai informação de cor através da média e variância aplicada em cada imagem. O BoC é um histograma de cor, cujo o objetivo é extrair a frequência de determinadas cores. O OpC é uma variante que aplica o BoW em cada canal de cor. A análise foi feita através do ANOVA (Análise de Variância) onde o valor-p < 2e-16 indicou uma diferença estatística entre as técnicas analisadas. A técnica OpC com o classificador SMO apresentou o maior desempenho, em torno de 95%. A métrica utilizada é a porcentagem dos classificados corretamente.

2 Trabalhos Correlatos

Em razão de ser uma metodologia de grande aplicação prática, a literatura apresenta algumas abordagens que utilizam a visão computacional para a contagem de leveduras e células.

Um trabalho que relacionou a fermentação do caldo de cana com leveduras e a contagem usando visão computacional foi o de Mongelo et al. [14]. Que avaliaram um software para efetuar a contagem e a identificação das leveduras viáveis e inviáveis por coloração com azul de metileno em imagens. Os autores tomaram amostras de caldo de cana em fermentação acompanhado o processo como ele ocorre na indústria, onde se inicia com

teor de açúcar de 12 Brix (120g/litro) em que o Brix representa a medida de concentração de açúcar e, à medida que o açúcar vai sendo consumido e o álcool vai sendo gerado. O Brix reduz e as leveduras perdem a viabilidade, situação ideal para acompanhar a viabilidade da levedura. As amostras foram coletadas nos pontos da fermentação equivalentes aos valores de Brix inicial (Brix 12,0), na metade da fermentação (Brix 6,0) e no final da fermentação (Brix 3,0), de modo a ter diminuição gradual de leveduras viáveis e aumento gradual de leveduras inviáveis. Para acompanhar a técnica foi utilizado o algoritmo K-curvatura em regiões de interesse (ROI) para a extração de atributos referentes à forma e a árvore de decisão para identificação. O banco de imagens utilizado foi construído através de imagens coletadas por microscópio. Os resultados obtidos pelos autores demonstraram que a aplicação identificou melhor as leveduras viáveis do que as inviáveis, isto porque a principal diferença é que as leveduras inviáveis apresentam à cor azul, quando misturadas com corante azul de metileno e a extração de atributos foi realizada por meio de forma das leveduras.

Schier et al. [11] utilizaram e avaliaram a técnica FRT (*Fast Radial Transform*) para contar colônias de leveduras em placas de Petri. Esta técnica foi utilizada para localizar leveduras na imagem, onde o maior problema foi contar as leveduras de forma separada, uma vez que cada colônia possui um grupo de leveduras que muitas vezes se apresentavam sobrepostas umas as outras. Os testes foram realizados com 245 imagens contendo colônias de leveduras de diferentes formas. Os resultados mostraram que a técnica FRT apresentou uma taxa de erro menor que 0.04 para a contagem de colônias de leveduras.

Coelho et al. [17] apresentaram uma análise de 8 técnicas para efetuar a segmentação com o intuito de localizar células em imagens coletadas de microscópio. As técnicas utilizadas foram: *AS Manual*, *RC Threshold*, *Otsu Threshold*, *Mean Threshold*, *Watershed* (*direct*), *Watershed* (*gradient*), *Active Masks* e *Merging Algorithm*. Dois bancos de imagens foram utilizados (U2OS e NIH3T3), totalizando em 4009 imagens microscópicas de células. Os autores realizaram uma segmentação manual (*AS Manual*) de 97 imagens microscópicas de células, para avaliar as segmentações resultantes das outras técnicas automáticas. Os

resultados obtidos demonstraram que a técnica *Merging Algorithm* apresentou os melhores resultados em todas as métricas que foram utilizadas para comparar o desempenho.

Entre as diferentes técnicas de extração de atributos, o algoritmo denominado Histograma de Palavras Visuais (*Bag-of-Word - BoW*) é uma técnica de extração de atributos resultante de um vetor de ocorrência de um vocabulário constituído por palavras visuais. Uma palavra visual corresponde a um ponto médio encontrado e descrito em uma imagem. Estes pontos são regiões localizadas por detectores de pontos de interesse, como por exemplo, o algoritmo SURF (*Speeded Up Robust Features*) [1].

Botterill et al. [19] descreveram uma aplicação para o reconhecimento de cenas e objetos em tempo real para auxiliar na navegação de robôs. As técnicas utilizadas foram o BoW em conjunto com o histograma de cor obtido do espaço de cor HSV. Os resultados obtidos foram relacionados ao tempo necessário para o reconhecimento, onde a taxa de tempo de reconhecimento se manteve em 0.0036 segundos por cada imagem, permitindo o reconhecimento em tempo real da cena.

3 Extratores de Atributos

Para efetuar a identificação de imagens é necessário extrair atributos que são utilizados na tarefa de distinguir e comparar imagens. As principais técnicas de visão computacional utilizadas neste trabalho para a extração de atributos são variações do algoritmo BoW (*Bag-of-Word*) que utilizam informação de cor como fator adicional, pois, o algoritmo BoW é aplicado somente em imagens em níveis de cinza.

3.1 Histograma de Palavras Visuais - BoW (*Bag-of-Word*)

O histograma de palavras (*Bag-of-Word*) [1] é um algoritmo utilizado no campo de reconhecimento de textos. Esta técnica extrai um histograma a partir da contagem das

ocorrências de palavras-chave em um texto. Palavras-chave é um conjunto de palavras utilizadas para distinguir um texto, de forma a identificar como, por exemplo, o tipo do texto: tecnológico, poético, romântico e etc. Os valores do histograma resultante do BoW são utilizados por classificadores para efetuar a identificação.

Assim como no reconhecimento de textos o algoritmo BoW passou a ser aplicado em imagens e denominado como Histograma de Palavras Visuais. Como o nome diz, as palavras passaram a ser visuais. Cada palavra visual corresponde um ponto que representa um conjunto de pontos de interesse com as mesmas características Csurka et al. [1].

O algoritmo BoW é composto pelos seguintes passos: detecção e descrição de pontos de interesse, criação do vocabulário e geração do histograma. A detecção e descrição de pontos de interesse correspondem à etapa de localizar e extrair informação de pontos, cuja variação em torno da vizinhança é maior em relação aos outros pontos. A criação do vocabulário permite utilizar pontos representantes de todo o conjunto de pontos de interesse localizados e descritos, para a geração de um histograma que corresponde à contagem de pontos de interesse que ocorrem em uma determinada imagem em relação ao vocabulário.

3.2 Algoritmo BoC(*Bag-of-Color*)

O algoritmo BoC é uma técnica de extração de atributos em imagens, também conhecida como extrator de assinatura de cor [4]. Este algoritmo pode extrair informação relacionada à cor de duas formas: global e local. Na forma global os atributos estão relacionados com toda a imagem. Já na forma local os atributos estão relacionados com as regiões capturadas por algoritmos de extração de pontos de interesse. Dado um conjunto de imagens, o algoritmo pode ser representado com os seguintes passos:

Algoritmo: BoC

1. Para cada imagem $\boldsymbol{i}$ de um conjunto de $\boldsymbol{N}$ imagens de treinamento, faça
2. Redimensione a imagem para *256x256* pixels e converta para o espaço de cor **CIELab**;
3. Divida a imagem em blocos de 16x16 pixels;
4. Para cada bloco $\boldsymbol{j}$, encontre a cor $\boldsymbol{c_{i,j}}$ de maior ocorrência, sendo os empates resolvidos aleatoriamente.
5. Aplique o algoritmo k-médias no conjunto de cores $\boldsymbol{c_{i,j}}$ e encontre as $\boldsymbol{k}$ cores mais representativas do conjunto de imagens de treinamento. As cores mais representativas correspondem aos centroides $\mathbf{R} = \{\mathbf{r_1}, ..., \mathbf{r_k}\}$;
6. Para calcular o histograma $\boldsymbol{h}$ de cada imagem $\boldsymbol{i}$, faça
7. Para cada pixel $\boldsymbol{p}$ da imagem $\boldsymbol{i}$, faça
8. Calcule a distância $\boldsymbol{d_{pi}}$ entre a cor do pixel $\boldsymbol{p}$ *e as cores representativas* $\boldsymbol{r_i}$ e incremente em uma unidade a posição do histograma que corresponde a cor mais próxima, isto é, a posição com a menor distância $\boldsymbol{d_{pi}}$ para $\boldsymbol{1 \leq i \leq k}$;
9. Normalize o histograma por meio da equação $\boldsymbol{h} = \frac{\boldsymbol{h}}{\sum_{\boldsymbol{j}=\mathbf{1}...\mathbf{k}} \boldsymbol{h_j}}$

3.3 Algoritmo CCV (*Color Coherence Vectors*)

O algoritmo CCV utiliza o conceito de coerência para a extração de atributos referentes à cor. A coerência refere-se ao grau de similaridade de cores apresentado em uma região de pixels. O grau de similaridade é definido como regiões que são representadas por uma mesma cor. O CCV obtém um histograma mais robusto em relação à informação espacial, que é um dos principais problemas encontrados em histogramas [2]. Na Figura 1 podemos visualizar o funcionamento do algoritmo CCV. O CCV foi aplicado na região em torno de cada ponto de interesse capturado por algum algoritmo detector de pontos de interesse, como por exemplo, o SURF.

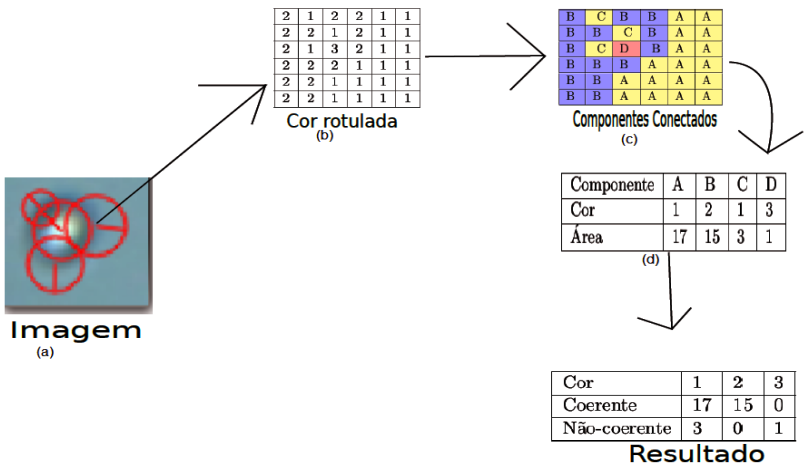


Figura 1: (a) Imagem com pontos de interesse detectados. (b) Uma região em torno de cada ponto de interesse com as cores rotuladas. (c) Detecção dos componentes conectados. (d) Contagem das regiões de cada ponto de interesse.

Os passos utilizados no CCV são os seguintes:

- Encontrar e rotular os componentes conectados.
- Efetuar a contagem de todos os pixels localizados em cada componente.
- Definir e aplicar um limiar para classificar a quantidade de elementos coerentes e não-coerentes para cada cor.

O CCV é aplicado em imagens no espaço RGB, como cada canal deste espaço utiliza 8 bits que resulta em mais de 16 milhões de cores, podemos reduzir. Para efeito de redução deste espaço utilizamos 2 bits para representar cada canal, o que resulta em 64 cores.

3.4 Algoritmo CM (*Color Moments*)

O algoritmo CM foi proposto por Bahrie Zouaki[4] como uma extensão do algoritmo SURF, denominada *SURF-Color Moments* (corresponde ao SURF + *Color Moments*). Esta extensão adiciona a informação de cor ao algoritmo SURF. A informação de cor é extraída ao redor de cada ponto de interesse localizado pelo SURF, através do cálculo da média e da variância dos valores dos pixels. Na Figura 2 podemos visualizar as matrizes resultantes da região em torno de cada ponto de interesse e que são utilizadas para calcular a média e a variância, representadas respectivamente pelas Equações 1 e 2. O E_i representa a média,

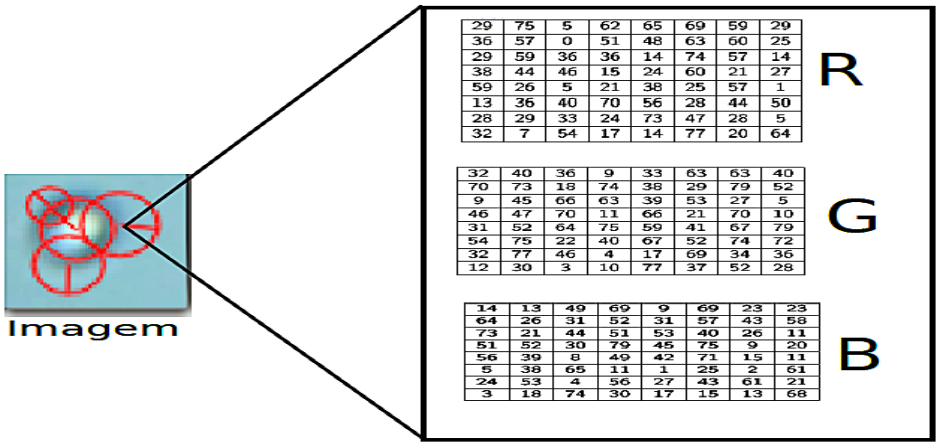


Figura 2: Imagem mostrando as matrizes com valores de uma região de um ponto de interesse. Cada matriz representa um canal da imagem.

onde i corresponde ao canal da imagem, N o tamanho da matriz e P_{ij} representa o pixel, σ_i representa a variância.

$$E_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij} \tag{1}$$

$$\sigma_i = \left[\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij} - E_i\right)^2\right]^{\frac{1}{2}} \tag{2}$$

Para cada ponto de interesse detectado pelo SURF, as estatísticas são calculadas em regiões de 5X5 segundo o artigo de Bahri e Zouaki [4]. Um vetor é preenchido com os resultados dos cálculos da média e da variância para cada canal da imagem, assim em cada ponto de interesse detectado na imagem no espaço RGB teremos um vetor de 6 posições. Este vetor é concatenado ao vetor que descreve o ponto de interesse.

3.5 Algoritmo OpC (*Opponent Color*)

Em Sande et al. [6] é abordada a utilização do SURF em imagens coloridas, com o objetivo de utilizar a informação de cor associada aos pontos de interesse detectados. A técnica OpC é uma extensão do algoritmo SURF, sendo a principal mudança relacionada com a matriz Hessiana (responsável pela detecção dos pontos de interesses). No Algoritmo SURF a matriz Hessiana é aplicada em imagens com um único canal, enquanto que, no OpC a matriz Hessiana é aplicada em imagens com três canais. O somatório do resultado de cada matriz Hessiana aplicada em cada canal é utilizado para encontrar os pontos de interesse.

A matriz Hessiana corresponde ao passo inicial para a detecção de pontos de interesse. O valor do determinante dessa matriz é utilizado como uma referência de um possível ponto de interesse, uma vez, que quanto maior for este valor, maior serão as mudanças entorno de cada ponto. Os pontos com os valores maiores que um valor adotado como limiar, serão definidos como pontos de interesse. Na Equação 3, podemos visualizar a

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \tag{3}$$

representação de uma matriz Hessiana, onde o determinante corresponde às variações em torno de cada ponto. $L_{xx}(X,\sigma)$ corresponde a convolução da derivada de segunda ordem de $\frac{\partial^2 g(\sigma)}{\partial x^2}$ no ponto $X = (x, y)$ e escala σ, sendo similar para $L_{yy}(X,\sigma)$ e $L_{xy}(X,\sigma)$.

4 Experimentos

4.1 Classificadores

Para identificação de viabilidade de leveduras é necessário classificar as imagens de leveduras após a extração de atributos. Foram utilizadas técnicas de classificadores supervisionados, ou seja, que pressupõem o conhecimento de uma base de treinamento com imagens de leveduras previamente classificadas.

Os classificadores selecionados foram Máquinas de Vetores de Suporte (SVM - *Support Vector Machines*) [6], Árvore de Decisão [10], NaiveBayes [10] e K-vizinhos mais próximos (KNN *K-Nearest Neighbor*) [6]. Estes classificadores foram utilizados no ambiente WEKA[1]. É importante ressaltar que o WEKA utiliza siglas para cada um desses classificadores, sendo eles: IBk para o KNN, J48 que é uma implementação do algoritmo C4.5 que é uma árvore de decisão, NBB para Naive Bayes e SMO que é uma implementação otimizada do SVM. Desta forma, de agora em diante utilizaremos as siglas utilizadas pelo WEKA para referenciar cada um dos classificadores utilizados.

[1] Site: http://www.cs.waikato.ac.nz/ml/weka/.

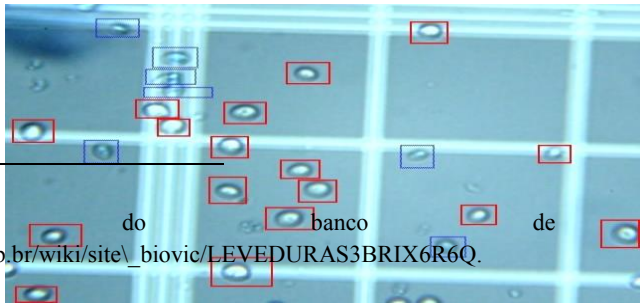
4.2 Validação Cruzada

Nos experimentos a técnica de validação cruzada com 10 partições foi adotada. A técnica Validação Cruzada permite distribuir um conjunto de dados em um número fixo de partições com quantidades aproximadamente iguais. Em geral n-1 partições são utilizadas para treinamento e uma partição para teste, sendo n o número de partições. A partição de teste é mudada a cada iteração de forma que, cada partição seja utilizada uma vez para teste.

4.3 Banco de Imagens

As imagens utilizadas neste trabalho foram extraídas do banco de imagens de leveduras [2]. Este banco foi construído pelo grupo INOVISÃO (Desenvolvimento e Inovação em Visão Computacional), que realizaram experimentos com fermentação. A fermentação foi obtida através da adição de leveduras *Sacchoromyces cerevisae* ao mosto na concentração de 1% (p/v). O mosto foi ajustado em 12 Brix sendo também utilizadas as amostragens quando o valor do Brix ficou em 6 e 3. O Brix corresponde à medida da concentração de açúcar no caldo da cana em suas diluições denominadas mosto.

Para efetuar a contagem a câmera de Neubauer foi utilizada e as amostras contendo as leveduras foram misturadas em corante azul de metileno, com o intuito de colorir com a cor azul as leveduras com baixa atividade fisiológica [13]. Na Figura 3 podemos observar uma imagem obtida da câmera de Neubauer.



[2] Local do banco de imagens: http://trac.gpec.ucdb.br/wiki/site_biovic/LEVEDURAS3BRIX6R6Q.

Figura 3: Imagem de leveduras observadas por microscópio usando câmara de NeuBauer. Imagem com leveduras viáveis (quadrado de cor vermelha), e leveduras inviáveis (quadrado de cor azul).

Foram recortadas 2614 imagens manualmente de 30 imagens obtidas no momento em que a fermentação se encontrava com Brix 3, e definidas 3 classes de imagens: inviável com 727 imagens, viável com 292 imagens e fundo com 1595 imagens. A classe viável corresponde às leveduras viáveis com alta atividade biológica e que são responsáveis pela fermentação e produção do etanol. A classe inviável corresponde às leveduras com baixa atividade biológica, já a classe fundo corresponde ao fundo da imagem, isto é, ausência de leveduras mas com linhas do retículo da câmara de Neubauer.

Na Figura 4 podemos visualizar exemplos de recortes, com leveduras viáveis, inviáveis e fundo com retículo.

Viável	Inviável	Fundo
		
		

Figura 4: Resultado de recortes de leveduras e fundo com retículo.

4.4 Métrica

A principal métrica utilizada para a comparação das técnicas foi a porcentagem de classificação correta. A porcentagem de classificação correta corresponde ao número de imagens identificadas corretamente de todas as classes, dividida pelo total de imagens.

4.5 Resultados e Discussões

Os resultados são apresentados por meio da combinação de classificadores e extratores. Os classificadores foram: KNN, Árvore de decisão, Naives Bayes e SVM. Os extratores de atributos foram: BoC, BoW, CCV, CM e OpC. Assim, a combinação entre o classificador IBk e o extrator de atributo BoC é referenciado como IBkBoC. As demais combinações são referencias de forma análoga.

A Figura 5 representa o diagrama de caixa obtido do software R[3]. Neste diagrama podemos visualizar o desempenho de cada combinação através da comparação das medianas, definidas pela faixa mais escura localizada em cada caixa. Observando o diagrama podemos verificar que a combinação SMOOpC apresentou a mediana com o maior desempenho em relação às outras medianas. Como descrita anteriormente, esta técnica corresponde ao classificador SMO com o extrator de atributos *Opponent Color*. Podemos observar alguns comportamentos incomuns, como as combinações do extrator de atributos BoC com os classificadores, quase que apresentam os melhores desempenhos, exceto com o classificador SMO. As combinações do classificador IBk com os extratores de atributos, apresentam quase que todos os piores resultados.As combinações de extratores de atributos com o classificador SMO, quase que apresentam os melhores resultados. As técnicas SMOOpC e J48BoC apresentaram os melhores resultados.

[3] Software estatístico, mais informações em: http://www.r-project.org/.

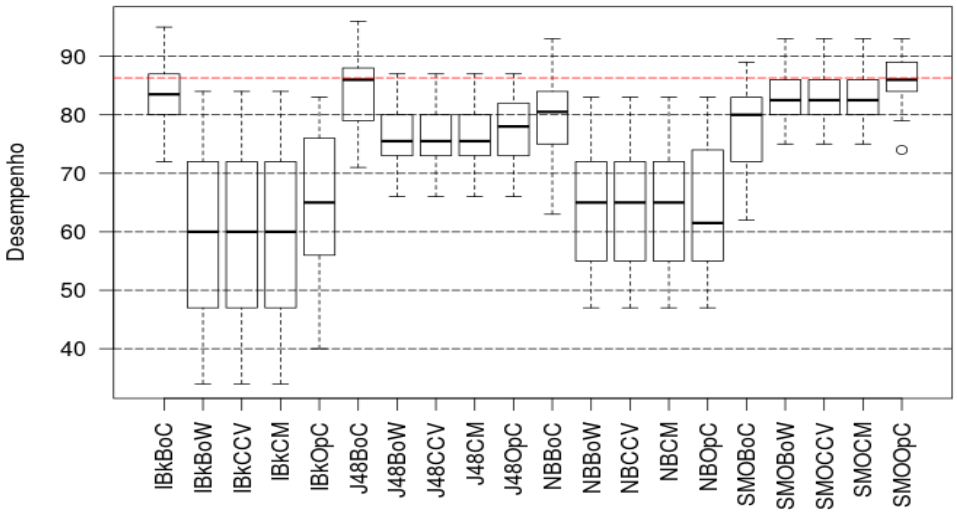


Figura 5: Diagrama de caixa das técnicas analisadas. Cada técnica corresponde ao classificador combinando um extrator de atributos.

4.5.1 ANOVA

Para validar os resultados e mostrar que as diferenças amostrais são significativas através de hipóteses, aplicamos a análise de variância (ANOVA) e obtivemos o valor-p < 2e-16. Este resultado indica que a hipótese nula pode ser descartada, ou seja, as medianas indicam que existe uma diferença estatística entre as técnicas.

4.5.2 Problemas encontrados

A técnica SMOOpC com o dicionário no tamanho 256 apresentou o melhor desempenho na identificação das imagens de leveduras incolores (viáveis). Para ilustrar um dos problemas de classificação encontrados no trabalho, tomamos como exemplo uma das 30 imagens disponíveis no banco de dados. Essa imagem tem 121 instâncias classificadas,

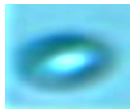
conforme a matriz de confusão apresentada na Figura 6. Pela matriz de confusão tem-se 82 instâncias identificadas como fundo e 16 instâncias identificadas como leveduras viáveis.

a	b	c	<- classe
82	1	12	| a= fundo
3	2	1	| b=inviável
4	0	16	| c=viável

Figura 6: Matriz de confusão.

Na Figura 7 podemos visualizar duas leveduras que foram classificadas como viáveis, porém como podemos observar, a identificação está incorreta. Este foi um dos problemas encontrados, onde as duas imagens foram confundidas, pois ambas as leveduras apresentam a mesma forma, porém a cor da região central de cada levedura é o fator determinante para classificá-las. Como não temos um controle sobre as regiões detectadas pelo algoritmo OpC então os pontos de interesse podem ser detectados em qualquer região da imagem. Isto significa que não temos a informação espacial, e este é um dos principais problemas que encontramos em histogramas.

Levedura viável.



Levedura Inviável.

Figura 7: Imagem a esquerda corresponde a levedura viável e imagem da direita levedura inviável. Ambas as imagens classificadas como viáveis.

Os resultados demonstraram que a informação de cor adicionada ao algoritmo BoW, melhorou os resultados na identificação das leveduras. Mesmo existindo alguns problemas na identificação, o algoritmo *Opponent Color* com o classificador SMO apresentou o melhor

desempenho com um dicionário de tamanho 256. Em relação ao trabalho de Mongelo et al. [14] o nosso resultado é melhor em relação à taxa de identificação de leveduras viáveis, o que é o ponto importante para verificar a eficiência da fermentação industrial.

5 Conclusão

Para garantir a qualidade na produção do etanol, as atividades de contagem e classificação de leveduras são cruciais. A contagem por diferença de coloração para viáveis e inviáveis atende às necessidades das usinas pela rapidez de execução, mas não pela exatidão dos resultados em função da dependência da visão humana. Uma forma de automatizar este processo é utilizar à visão computacional. Com isto, neste trabalho analisamos o algoritmo BoW em conjunto com algumas técnicas que capturam a informação referente a cor, para extrair atributos que foram utilizados em classificadores. Os resultados obtidos demonstraram que a técnica de extração de atributos *Opponent Color* com o classificador SMO obteve os melhores resultados. A métrica utilizada foi a porcentagem de classificados corretamente. Com o teste de hipótese ANOVA e o valor-p < 2e-16 observou-se diferença estatística entre as técnicas analisadas. Com a matriz de confusão verificamos que a técnica SMOOpC com o dicionário de tamanho 256 identificou melhor as leveduras viáveis e o fundo, mesmo apresentando erros na identificação das leveduras inviáveis, a técnica SMOOpC apresenta um desempenho melhor em relação as outras técnicas analisadas.

Bibliografia

[1] Csurka, G. et al. ***Visual categorization with bags of keypoints***. Workshop on statistical learning in computer vision, *ECCV*, Vol. 1, 1-22. Nº 1, 2004.

[2] Pass, G.; Ramin, Z.; Miller, J. ***Comparing images using color coherence vectors***. Proceedings of the fourth ACM international conference on Multimedia, ACM, p. 65-7, 1997.

[3] Bahri, A.; Zouaki, H. ***A surf-color moments for images retrieval based on Bag-of-Feature.*** European Journal of Computer Science and Information Technology, v. 1, p. 11-22, 2013.

[4] Wengert, C.; Douze, M.; Jégou, H. ***Bag-of-colors for improved image search.*** Proceedings of the 19th ACM international conference on Multimedia, ACM, p. 1437-1440, 2011.

[5] Sande, K. V.; Gevers, T.; Snoek, C. G. ***Color descriptors for object category recognition.*** Conference on Colour in Graphics, Imaging and Vision, Society for Imaging Science and Technology,Vol. 2008, p. 378-381, Nº. 1, 2008.

[6] Witten, I. H.; Frank, E. ***Data Mining: Practical machine learning tools and techniques.*** Morgan Kaufmann, San Francisco, 2005.

[7] Weijer, J. V.; Khan, F. S. ***Fusing color and shape for bag-of-words based object recognition.*** Computational Color Imaging, Springer, p. 25-34, 2013.

[8] Stratford, M. ***Yeast flocculation: reestruturing the theories in line with recent research.*** Cerevisia, p. 38-45, 1996.

[9] Souza, S. C. ***Avaliação da produção de etanol em temperaturas elevadas por uma linhagem de Saccharomyces cerevisiae***. Tese (Doutorado em Biotecnologia), São Paulo, n. USP/Instituto Butantan/IPT, p. 49, 2009.

[10] Silva, D. S. et al. ***Classificação de Leveduras Utilizando Transformada de Hough e Aprendizagem Supervisionada***. Workshop de Visão Computacional 2012, Goiania - GO: In: WVC, 2012.

[11] Schier, J. K.; , B. ***Automated Counting of Yeast Colonies using the Fast Radial Transform Algorithm.*** Bioinformatics, p. 22-27, 2011.

[12] Salama, G.; Abdelhalim, M.; Zeid, M. A.-E. ***Breast Cancer Diagnosis on Three Different Datasets using Multi-classifiers.*** Breast Cancer (WDBC), v. 32, p. 2, ISSN 569, 2012.

[13] Nara, L. Q. et al. ***Classificação de Leveduras para o Controle Microbiano em Processos de Produção de Etanol.*** Presidente Prudente - SP: VI Workshop de Visão Computacional, 2010.

[14] Mongelo, A. I. et al. ***Validação de Método baseado em Visão Computacional para automação de Contagem deViabilidade de Leveduras em Indústrias Alcooleiras***. Bento Gonçalves - RS: In: VIII Congresso Brasileiro de Agroinformática SBIAGRO 2011, p. 17-21, 2011.

[15] Lima, U. A. ***Biotecnologia Industrial Processos Fermentativos e Enzimáticos***. São Paulo: Edgard Blücher, v. 3, p. 125-154, 2001.

[16] Domingues, A. T. ***O Setor Agroindustrial Canavieiro no Mato Grosso Do Sul: Desdobramentos e Perspectivas.*** Revista Tamoios, p. 21-36, 2012.

[17] Coelho, L. P.; Shariff, A.; Murphy, R. F. ***Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms.*** Biomedical Imaging: From Nano to Macro 2009, ISBI'09, IEEE International Symposium on, p. 518-521, 2009.

[18] Ceccato-Antonini, S.R.***Microbiologia da fermentação alcoólica: a importância do monitoramento microbiológico em destilarias***. São Carlos: EdUFSCar, Vol. 27, p. 120, Nº 5, 2010.

[19] Botterill, T.; Mills, S.; Green, R. ***Speeded-up bag-of-words algorithm for robot localisation through scene recognition*** .Image and Vision Computing New Zealand 2008, *IVCNZ 2008. 23rd International Conference,* Image and Vision Computing New Zealand, 2008, IVCNZ 2008. 23rd International Conference, p. 1-6, 2008.

[20] Borzani, W. et al. ***Biotecnologia Industrial***. Edgar Blucher LTDA, São Paulo, v. 1, 2001.

[21] Boinot, F. ***Melle Process of Alcoholic Fermentation. Usines de Melle***. The International Sugar Journal, France, p. 466-467, 1939.