

Evaluation of Machine Learning and Bag of Visual Words Techniques for Pollen Grains Classification

Carolini Nascimento Martins Rodrigues, Ariadne Barbosa Gonçalves, Gercina Gonçalves da Silva, Hemerson Pistori

Universidade Católica Dom Bosco – UCDB
Avenida Tamandare, 6000. Jardim Seminario. Campo Grande -MS

E-mail: { [carolinirodrigues28](mailto:carolinirodrigues28@gmail.com), [ariadne.gon](mailto:ariadne.gon@gmail.com), [gercina.goncalves](mailto:gercina.goncalves@gmail.com) } @gmail.com; pistori@ucdb.br

Abstract— *This work presents some new results regarding the automation of pollen grains classification using computer vision. A technique based on the Bag of Visual Words and Supervised Machine Learning algorithms is proposed and evaluated. A dataset of pollen grain images taken from 9 different pollen types was created and used in the experiments.*

Keywords— *Histograma de Palavras Visuais, Aprendizagem Supervisionada, Pólen.*

I INTRODUÇÃO

A identificação da origem do mel e outros produtos apícolas, como o pólen e a própolis, pode ajudar no desenvolvimento da produtividade apícola, bem como impulsionar a economia da região. Na produção de produtos apícolas as abelhas operárias de *Apis mellifera* saem da colmeia e percorrem um raio de voo de 3 km [18] a procura de flores recursos florais. Nesse processo, grãos de pólen ficam aderidos à abelha, e são carregados para dentro da colmeia. Assim, ao analisar amostras de produtos apícolas é possível identificar os grãos de pólen presentes neles, pois o pólen é uma marca indelével da procedência botânica do produto, o que influencia diretamente na sua denominação de origem.

A determinação da origem botânica de produtos apícolas é feita pela classificação visual dos tipos polínicos presentes no produto, sendo que ela pode ser muito morosa e cansativa, passível de erros devido ao desgaste físico e mental do especialista. Ressalta-se que a área que realiza o estudo dos grãos de pólen é a Palinologia. O bioma Cerrado possui uma flora muito rica, e consequentemente com grande variabilidade na morfologia dos polens. No entanto, alguns polens podem ser muito semelhantes o que dificulta a classificação e requer acurácia do especialista nesta tarefa. Além disso, existem semelhanças intraespecíficas dificultando a classificação tornando a automação da identificação polínica por visão computacional uma maneira precisa e rápida de classificação dos grãos de pólen.

A técnica de Histograma de Palavras Visuais ou Bag of Words (BOW) refere-se a um vetor de atributos que é extraído de determinada imagem ou conjunto de imagens. Para a extração do atributo é necessário o uso de algoritmos auxiliares como os de detecção de pontos de interesse SIFT ou SURF, e em conjunto com o BOW são utilizados algoritmos de aprendizagem automática. O objetivo do trabalho é a

utilização da técnica BOW para classificação automática de grãos de pólen.

O trabalho está organizado em 6 seções. Na Seção II será apresentada a base teórica referente a trabalhos correlatos utilizando o BOW e de trabalhos com classificação e contagem de grãos de pólen. Na Seção III é descrita a técnica e as utilidades do BOW. A Seção IV apresenta a metodologia utilizada para o experimento. A Seção V apresenta os resultados e discussões sobre os experimentos, e as conclusões em conjunto com propostas para trabalhos futuros estão expostos na Seção VI.

II TRABALHOS CORRELATOS

O problema da identificação de padrões visuais em coleções de imagens na histologia é abordado em [5]. O método proposto representa um conteúdo visual usando o BOW. O método consiste em um mecanismo de interpretação de dados associados às regiões das imagens correspondentes, para tanto dois conjuntos de imagens foram criados. O primeiro com imagens histológicas, e o segundo foi composto por imagens microscópicas usadas para diagnósticos.

Diferentes representações de palavras visuais e codebooks [10] foram testadas, e o desempenho foi avaliado qualitativamente e quantitativamente. O método explora a representação do BOW para realizar a mineração de dados visuais a partir de uma perspectiva mais ampla, em que o foco seja o conjunto de imagens e não as imagens individuais. Os resultados demonstraram que o BOW é uma boa alternativa para representar o conteúdo visual de imagens histológicas.

A abordagem do BOW para detecção de melanoma maligno com base em imagens microscópicas de epiluminescência foi aplicada em [15]. Os algoritmos utilizados para classificação foram o SVM e o Naive Bayes. O conjunto de 100 imagens utilizadas foram de pele humana lesionada, representadas através de histogramas. Através deste conjunto de imagens os algoritmos foram testados e treinados, sendo 70 imagens com lesões provocadas por motivos adversos e 30 imagens com lesões provenientes do melanoma. Nas imagens foram aplicados filtros, como o Wavelet e o "Gabor-like". Os resultados indicam que o algoritmo SVM teve melhor desempenho em relação ao conjunto de imagens.

Um protótipo de um sistema para classificação automática de imagens de câncer colorretal captadas pela técnica NBI (Imagiologia de banda estreita) foi apresentado em [17]. O sistema consiste em uma representação BOW, com o uso do detector de ponto de interesse SIFT e o classificador

SVM. Os experimentos foram feitos em um conjunto de 907 imagens de tumores colorretais, sendo que o tamanho do tumor varia entre 100 x 300 a 900 x 800 e nesse conjunto há imagens que compartilham o mesmo tumor. As imagens foram divididas em 10 grupos, onde 9 foram usadas para treinamento e o restante para teste. A classificação obteve uma taxa de reconhecimento máximo de 94,1%.

No trabalho [7], é feita a classificação de três tipos diferentes de grãos de pólen pertencentes à família Urticaceae. Devido à reação alérgica no ser humano, as práticas clínicas mostram grande interesse na diferenciação desses grãos de pólen. Informações morfológicas são utilizadas por especialistas para distinguir diferentes tipos de grãos de pólen. O objetivo do trabalho de [7] foi de desenvolver um sistema automático de reconhecimento de padrões para classificar os grãos de pólen. Nas imagens foram aplicadas as operações de erosão e dilatação para obter o tamanho do grão de pólen. O trabalho concentra-se em um conjunto de formas baseadas em características de modo a conseguir a classificação. Um conjunto de trinta imagens foi usado com 10 imagens de cada tipo de grão de pólen. Uma vez que todos os atributos foram calculados, os dados foram passados para um classificador de distância mínima. A distância utilizada foi a de Mahalanobis para medir a similaridade entre um caso de consulta e as classes do conjunto de imagens. Se a imagem é corretamente classificada soma um ao contador. O processo repete-se N vezes, cada vez com exclusão de uma imagem diferente. Obteve-se uma percentagem de classificação correta de 97%.

A classificação automática de grãos de pólen foi proposta em [8] onde são utilizados dois métodos para classificação. O primeiro é o modelo de contorno ativo (snake), e o segundo é o Paradise Network (rede paraíso). O pólen tem um limite de espessura, ou de dois gumes quando visto através de um microscópio óptico, logo o modelo snake é usado para detectar esses dois gumes, se eles estiverem presentes.

Uma análise de grãos de pólen foi realizada por [4] com o objetivo de contar o número de pólen nas imagens. Foi aplicada a técnica de watershed para o processamento da imagem e em seguida o detector de partículas para realizar a contagem. O resultado foi similar à contagem humana. O trabalho concluiu a possibilidade de obter uma alternativa eficaz para a contagem de grão de pólen.

III HISTOGRAMA DE PALAVRAS VISUAIS

A técnica de Histograma de Palavras Visuais (*Bag of Visual Words* – BOW) é composta de um histograma com números dos descritores de padrões encontrados em uma imagem. Para a classificação de imagens é aplicado o mesmo procedimento. Contudo a definição do vocabulário é um procedimento um pouco mais complexo. São usadas técnicas auxiliares denominadas SIFT e SURF [2]. Essas técnicas definem pontos de interesse na imagem (Figura 1). Os pontos são descritos por valores numéricos, os vetores. A quantidade de pontos em cada imagem é variável, o que impossibilita a utilização da técnica por algoritmos de aprendizagem baseados em vetores de atributos, por isso o BOW utiliza um algoritmo de agrupamento chamado K-médias (K-Means) em todos os pontos de interesse da imagem agrupados, e para cada grupo é calculada uma média. As médias ou número de grupos do K-

Médias definem a quantidade de palavras visuais do vocabulário em que cada grupo representa uma palavra. Após esse processo as médias também são utilizadas na extração de atributos de cada imagem.

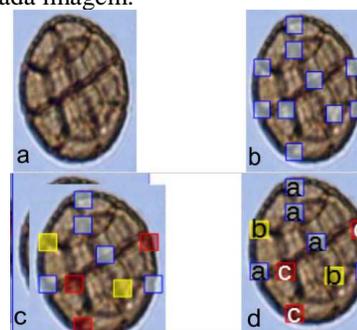


Figura 1: Atributos extraídos pelo BOW. a) Imagem original; b) detecção dos pontos de interesse pelo algoritmo Surf; c) Agrupamento dos vetores pelo K-means. d) Cada agrupamento é transformado em uma palavra pelo BOW.

O conjunto de treinamento é formado por imagens referentes ao problema, divididas em classes. Por exemplo, na classificação de pólen, as imagens são divididas de acordo com cada espécie. Os pontos de interesse das imagens são associados a cada palavra visual, de acordo com a menor distância entre o ponto analisado e a média de cada grupo. É criado um histograma com a quantidade de pontos de interesse de cada grupo. A frequência do histograma é usada como característica da imagem, que é associada a sua respectiva classe. Ao analisar uma nova imagem o processo é o mesmo da análise de pontos de interesse das imagens que são associados a cada palavra visual, de acordo com a menor distância entre o ponto analisado e a média de cada grupo.

IV MATERIAIS E MÉTODOS

Neste trabalho foram analisados somente os méis florais, pois possuem diversidade de grãos de pólen. As amostras de mel necessitam de processamento para extração dos grãos de pólen que estão contidos no mel, para que seja possível a sua observação em microscópio, pois não é possível a visualização dos detalhes do grão de pólen a olho nu. Para o isolamento dos grãos de pólen nas amostras de mel foi utilizado o método de acetólise [11]. Para tanto é necessário pesar uma alíquota de 10 g de mel, que foi diluído em 10 ml de água destilada morna, e adicionar 50 ml de álcool etílico 95% em cada amostra. A mistura foi centrifugada a 2500 RPM para que os grãos de pólen fossem isolados, sendo que na parte superior do recipiente ficou um material líquido (sobrenadante) que foi descartado, e o sedimento polínico que é o material que contém a concentração de grãos de pólen ficou aderido no fundo do recipiente (decantação). O sedimento recebeu 2 ml de ácido acético para retirada da água excedente da amostra e então foram acrescentados 2 ml da solução acetolítica. Os tubos foram colocados em banho-maria, para que ocorresse a reação da solução acetolítica com o calor, pois este processo retira o conteúdo líquido do interior do grão de pólen, deixando apenas o esqueleto do pólen. Após este processo, o sobrenadante foi descartado e 2 ml de glicerina foram adicionadas à amostra, que ficou em repouso por uma hora, pois este processo reidrata o pólen fazendo com que retorne a sua forma original. Uma demonstração do uso do processo de acetólise pode ser visto na Figura 2, que apresenta imagem do

tipo *Protium*. As lâminas polínicas foram preparadas da seguinte maneira: um cubo de gelatina glicerinada com dimensão 3 mm por 3 mm foi cortado e posicionado no fundo do recipiente para que os grãos de pólen pudessem se aderir ao cubo. O cubo foi colocado em uma placa de vidro de formato retangular (lâmina e lamínula) para então ser analisado no microscópio. As imagens dos tipos polínicos foram capturadas através de um microscópio digital que possui uma Monotela LCD, da marca LCD Micro Bresse com aumento de até 40X. Foi feita a captura das imagens, de diferentes ângulos dos grãos de pólen.

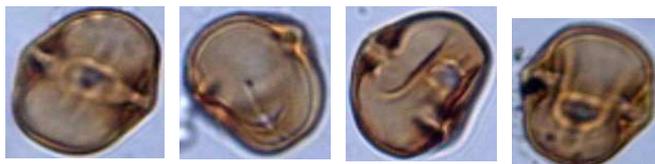


Figura 2. Imagens do tipo *Protium* obtidas através de acetólise.

Foram selecionados nove tipos polínicos (Figura 3) mais comuns em alguns méis do Estado de Mato Grosso do Sul e de cada tipo foram capturadas 30 imagens, totalizando um banco com 180 imagens, denominado de POLEN9E. As imagens contidas no banco de imagens foram utilizadas pela técnica BOW que identifica a variação de gradiente que ocorrem na imagem. O tamanho do dicionário utilizado pelo BOW representa o número de atributos e características que são extraídas do conjunto de imagens, para gerar o vetor de atributo Foram feitos testes exponenciais de 2 à 16384 no tamanho do dicionário, para saber o melhor tamanho de vocabulário. Experimentos foram realizados para verificar qual o melhor algoritmo de aprendizagem de máquina e o melhor tamanho de dicionário para melhor desempenho das informações extraídas do conjunto de imagens.

Foram verificados quatro algoritmos de aprendizagem, com o objetivo de obter um algoritmo ideal para este conjunto de imagens, foram testados: Máquina de Vetores de Suporte (SVM), C4.5, K Vizinhos Mais Próximos (KNN) e AdaBoost (Adaptive Boosting) ou estímulo adaptativo. Foram realizados alguns experimentos para testar o algoritmo KNN para escolher o melhor parâmetro para k a ser utilizado.

Os algoritmos foram comparados através das métricas: área sob a curva ROC (Area Under Roc), tempo de uso da CPU para teste (User CPU Time Testing), porcentagem de acerto (Percent Correct) e média ponderada da medida F (Weighted Average F Measure). Todos os experimentos foram feitos utilizando a validação cruzada com 10 dobras, presente no software WEKA.

Para teste de hipótese, utilizou-se o teste T-Student. A matriz de confusão foi utilizada para uma melhor visualização dos resultados dos experimentos, apresentando o resultado da classificação realizada pelo algoritmo. Cada linha representa uma espécie de pólen e cada coluna informa a quantidade de imagens que foram classificadas de acordo com a espécie. A diagonal da matriz demonstra a quantidade de imagens classificadas corretamente. Para a produção da matriz confusão, que apresenta os resultados obtidos pelo classificador com melhor desempenho, também foi utilizada a

validação cruzada com 10 dobras presentes no programa WEKA.

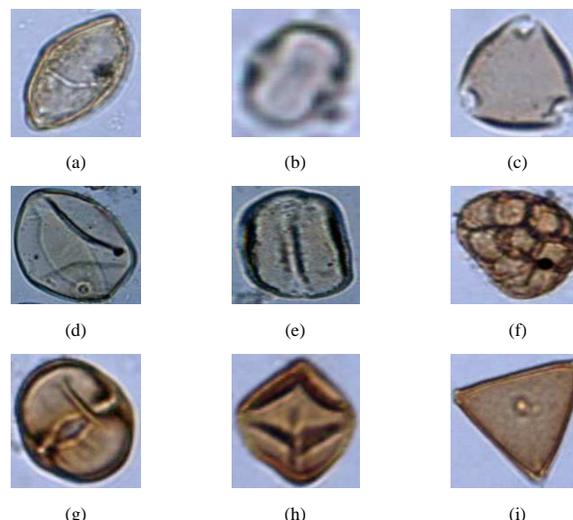


Figura 3. Imagens dos nove tipos polínicos presentes no banco de imagens. Arecaceae (a), *Cecropia* (b), *Myrcia* (c), Poaceae (d), *Syagrus* (e), *Anadenanthera* (f), *Protium* (g), *Schinus* (h), *Serjania* (i).

V RESULTADOS E DISCUSSÕES

Com os experimentos realizados foi possível uma análise do desempenho dos algoritmos de aprendizagem para este tipo de problema. Nas tabelas 1 à 4 são demonstrados os resultados obtidos com o experimento. Na primeira coluna da esquerda estão os tamanhos de dicionário e na segunda coluna o primeiro algoritmo a ser comparado, e nas colunas seguintes da direita estão os demais algoritmos de aprendizagem. Os experimentos com o KNN resultaram em um $k = 15$, pois obteve um bom resultado em todas as comparações feitas.

A Tabela 1 apresenta os resultados da métrica Porcentagem de acerto com o algoritmo SVM sendo comparado com os demais algoritmos. A ordem dos algoritmos ordem interfere de forma direta na análise do desempenho dos algoritmos, pois o primeiro é comparado com os demais algoritmos que apresentam seus desempenhos melhores ou piores em relação ao primeiro. Foram obtidas outras tabelas como esta, porém com variação da ordem dos algoritmos, contudo mantendo esta métrica. Fazendo uma análise de forma geral de todos os resultados, é possível obter alguns dados que levam a visualizar para esta métrica um algoritmo ou conjunto de algoritmos mais eficiente.

A primeira análise foi em relação à métrica Porcentagem de acerto e feito esse estudo foi possível observar um índice maior de alguns algoritmos como os de melhores desempenhos. O classificador SVM apresentou melhor desempenho nos dicionários de tamanho 256 até 16384, em dicionários de tamanho inferior teve um resultado relativamente igual aos outros classificadores e com base nesses resultados é possível afirmar que o SVM tem melhor desempenho nesta métrica.

TABELA 1: Porcentagem de acerto

Dic.	SVM	KNN	C4.5	AdaBoost
Dic02	14.27066	14.27920	12.82051	12.76781
Dic04	23.34900	23.10826	27.58262	15.82194 ^o
Dic08	23.65527	20.84046	29.99430 ^o	20.30342
Dic16	23.68234	23.51140	33.43305 ^o	19.46439
Dic32	29.12963	27.10399	40.73219 ^o	24.20370
Dic64	43.60541	37.81766	41.47436	23.60399 ^o
Dic128	50.79060	47.39744	45.28348	23.15527 ^o
Dic256	65.36182	50.13533 ^o	44.77208 ^o	23.30057 ^o
Dic512	64.71429	29.46746 ^o	37.25952 ^o	16.82540 ^o
Dic1024	68.93162	20.28632 ^o	35.80199 ^o	21.38319 ^o
Dic2048	74.03846	13.52564 ^o	36.75926 ^o	17.69231 ^o
Dic4096	71.24644	13.52564 ^o	38.86325 ^o	19.31197 ^o
Dic8192	64.82479	13.52564 ^o	37.03989 ^o	16.27635 ^o
Dic16384	59.35755	13.52564 ^o	36.02991 ^o	14.72365 ^o

^o, ^odegradação ou melhoria estatisticamente significativa com nível de confiança de 95%.

Nesta seção será verificada a tabela com o SVM como primeiro algoritmo dos quatro. Na Tabela 2 estão os resultados obtidos com a métrica Área sob a curva ROC, em que é possível observar os dados em relação ao desempenho de cada algoritmo, e analisar a frequência dos algoritmos que apresentaram melhores desempenhos que os outros.

TABELA 2: Área sob a curva roc.

Dic.	SVM	KNN	C4.5	AdaBoost
Dic02	0.51232	0.54460	0.50884	0.57671
Dic04	0.88855	0.81285	0.78349	0.52604 ^o
Dic08	0.72112	0.72522	0.91464	0.80386
Dic16	0.81743	0.79911	0.90138	0.60014 ^o
Dic32	0.58646	0.53383	0.88178 ^o	0.60597
Dic64	0.91388	0.87948	0.89099	0.59530 ^o
Dic128	0.93886	0.91418	0.83312 ^o	0.62248 ^o
Dic256	0.99512	0.99667	0.82828 ^o	0.64099 ^o
Dic512	0.99132	0.95389	0.70877 ^o	0.53602 ^o
Dic1024	0.99775	0.81863 ^o	0.78275 ^o	0.60642 ^o
Dic2048	0.99594	0.54729 ^o	0.74811 ^o	0.58431 ^o
Dic4096	0.99707	0.50000 ^o	0.75321 ^o	0.56020 ^o
Dic8192	0.98238	0.50000 ^o	0.76420 ^o	0.51347 ^o
Dic16384	0.98325	0.50000 ^o	0.65010 ^o	0.52882 ^o

^o, ^odegradação ou melhoria estatisticamente significativa.

Os tamanhos de dicionário de 1024 a 16384 apresentam melhores desempenho para o classificador SVM, enquanto que para os outros tamanhos de dicionário este classificador obteve resultados semelhantes aos dos outros classificadores. A métrica Área sob a curva ROC teve o algoritmo SVM com uma maior incidência de vezes como melhor algoritmo.

O terceiro experimento (Tabela 3) foi feito utilizando a métrica Média Ponderada da medida F, sendo que para esta métrica o melhor resultado foi obtido com o algoritmo SVM. O algoritmo apresentou melhor desempenho nos dicionários de tamanho 256 a 16384, nos demais dicionários teve um desempenho relativamente parecido com os outros classificadores.

TABELA 3: Média Ponderada da medida F.

Dic.	SVM	KNN	C4.5	AdaBoost
Dic02	14.27066	14.27920	12.82051	12.76781
Dic04	23.34900	23.10826	27.58262	15.82194 ^o
Dic08	23.65527	20.84046	29.99430 ^o	20.30342
Dic16	23.68234	23.51140	33.43305 ^o	19.46439
Dic32	29.12963	27.10399	40.73219 ^o	24.20370
Dic64	43.60541	37.81766	41.47436	23.60399 ^o
Dic128	50.79060	47.39744	45.28348	23.15527 ^o
Dic256	65.36182	50.13533 ^o	44.77208 ^o	23.30057 ^o
Dic512	64.71429	29.46746 ^o	37.25952 ^o	16.82540 ^o
Dic1024	68.93162	20.28632 ^o	35.80199 ^o	21.38319 ^o
Dic2048	74.03846	13.52564 ^o	36.75926 ^o	17.69231 ^o
Dic4096	71.24644	13.52564 ^o	38.86325 ^o	19.31197 ^o
Dic8192	64.82479	13.52564 ^o	37.03989 ^o	16.27635 ^o
Dic16384	59.35755	13.52564 ^o	36.02991 ^o	14.72365 ^o

^o, ^odegradação ou melhoria estatisticamente significativa.

A Tabela 4 apresenta como primeiro algoritmo o C4.5, pois este apresentou melhor resultado com esta métrica. Com esta métrica obtivemos resultados parecidos entre alguns classificadores.

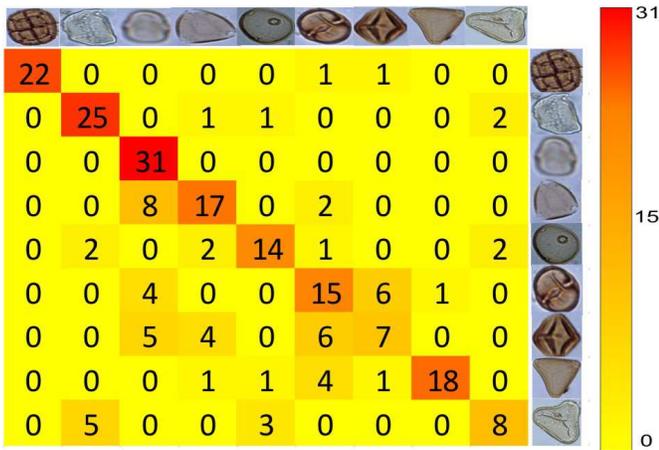
TABELA 4: Teste do tempo de uso da CPU.

Dic.	C4.5	SVM	KNN	AdaBoost
Dic02	0.00030	0.00030	0.00130	0.00010
Dic04	0.00000	0.00010	0.00160	0.00000
Dic08	0.00020	0.00020	0.00290	0.00000
Dic16	0.00000	0.00030	0.00530 ^o	0.00000
Dic32	0.00020	0.00030	0.01040 ^o	0.00020
Dic64	0.00010	0.00140	0.02190 ^o	0.00000
Dic128	0.00030	0.00260	0.04300 ^o	0.00000
Dic256	0.00010	0.00440 ^o	0.08810 ^o	0.00010
Dic512	0.00010	0.02750 ^o	0.31040 ^o	0.00030
Dic1024	0.00100	0.01920 ^o	0.37080 ^o	0.00040
Dic2048	0.00090	0.04190 ^o	0.71740 ^o	0.00060
Dic4096	0.00240	0.08050 ^o	1.42130 ^o	0.00100
Dic8192	0.00200	0.15860 ^o	2.91610 ^o	0.00140
Dic16384	0.00420	0.28960 ^o	6.39040 ^o	0.00390

^o, ^odegradação ou melhoria estatisticamente significativa.

Através da observação das quatro tabelas apresentadas nesta seção, é possível determinar que o algoritmo C4.5 tem um desempenho mais rápido do que os demais algoritmos. No entanto, apresenta um resultado muito parecido com o algoritmo AdaBoost. Assim apresenta-se o melhor resultado dos quatro algoritmos testados com esta métrica. A métrica Tempo de uso da CPU para teste faz uma análise do tempo que leva para o algoritmo testar o conjunto de dados gerado a partir das imagens. Como estes dois algoritmos obtiveram resultados parecidos é preciso analisar outros aspectos para verificar um melhor desempenho.

Com esses experimentos foi possível verificar que o algoritmo SVM teve uma boa classificação em relação aos outros classificadores. Por isso foi gerada uma matriz de confusão com o resultado de classificação deste algoritmo, para que este seja melhor avaliado. A matriz de confusão (Figura 4) a seguir apresenta os resultados da classificação do algoritmo de aprendizagem SVM.



a- *Anadenanthera*. b- *Arecaceae*; c- *Cecropia*; d- *Myrcia*; e- *Poaceae*; f- *Protium*; g- *Schinus*. h- *Serjania*; i- *Syagrus*.

Figura 4: Matriz de confusão gerada com o classificador SVM.

Para identificar visualmente quais imagens estavam sendo classificadas, outra classificação foi realizada a partir de um conjunto de 45 imagens de grãos de pólen. A Figura 5 mostra os resultados da classificação para o tipo *Arecaceae*. Foram classificadas corretamente as imagens (b), (d), (f) e (g). A imagem (a) e (c) são do tipo *Syagrus*, e a imagem (e) é do tipo *Poaceae*. A classificação errônea pode ter ocorrido devido às características parecidas, como por exemplo, a coloração e a textura.

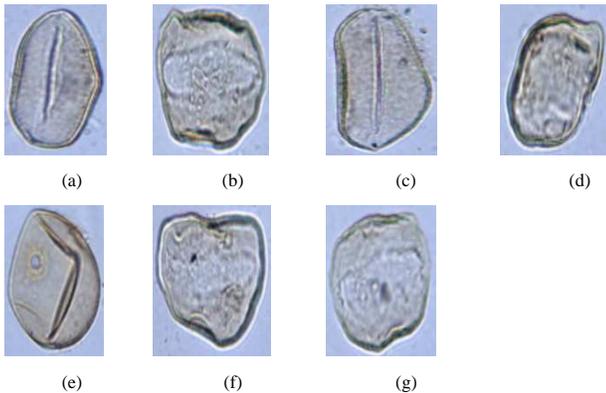


Figura 5. Imagens classificadas como *Arecaceae*.

A Figura 6 mostra os resultados do tipo classificado como *Cecropia*. Nesses resultados as imagens (a), (b), (d), (e) e (f) foram classificadas corretamente. As imagens (c) e (g) são do tipo *Myrcia*. O erro na classificação nesse caso pode ter sido decorrente do tratamento da imagem, pois a resolução de ambas as espécies não estão boas, o que pode causar certa confusão na hora da classificação.

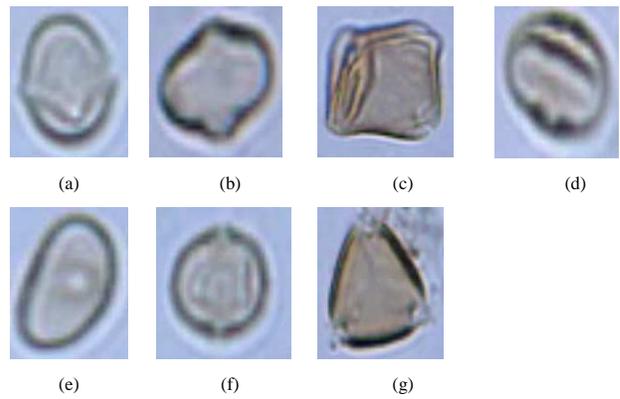


Figura 6. Imagens classificadas como *Cecropia*.

A Figura 7 apresenta as imagens resultantes da classificação para o tipo *Myrcia*. As imagens (b), (e) e (f) foram classificadas corretamente. A imagem (a) é de *Protium*, a imagem (c) é de *Arecaceae* e a imagem (d) é do tipo *Schinus*. Um atributo que podemos notar certa semelhança é em relação às formas de cada grão, também é possível observar que a cor se assemelha com a do tipo *Arecaceae*.

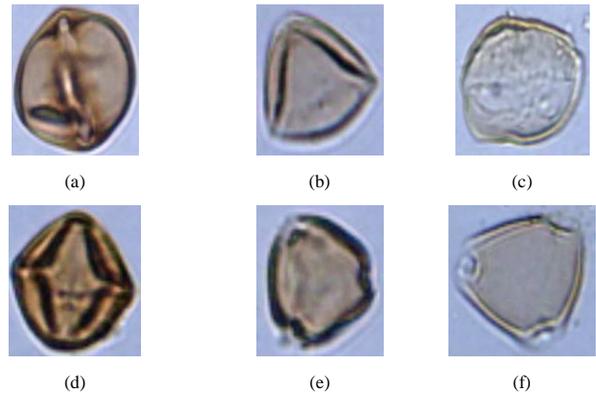


Figura 7. Imagens classificadas como *Myrcia*.

A Figura 8 mostra as imagens classificadas como *Poaceae*. As imagens (b), (c) e (d) estão classificadas corretamente. A imagem (a) é do tipo *Syagrus*. A textura de *Syagrus* é semelhante ao tipo *Poaceae*, como também sua forma o que fez com que o classificador classificasse essa imagem de *Syagrus* como *Poaceae*.

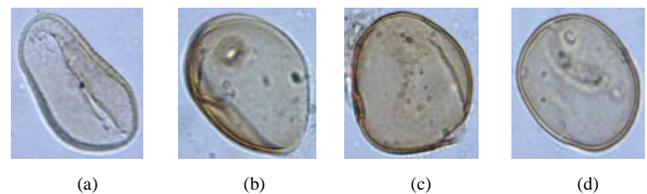


Figura 8. Imagens classificadas como *Poaceae*.

As imagens da Figura 9 apresentam os resultados da classificação para o tipo *Protium*, sendo que as imagens (b) e (d) foram classificadas corretamente. A imagem (a) é do tipo *Serjania* e a imagem (c) é do tipo *Schinus*. As imagens (b), (c) e (d) possuem formas muito semelhantes. A imagem (a) apresenta os atributos de cor e textura semelhante com a imagem (d). Analisando as semelhanças das imagens (d) e (a) podemos ver que a imagem (b), que é do mesmo tipo da

imagem (d), e não possuem semelhanças de cor e de morfologia com a imagem (a). Desta maneira podemos dizer que as imagens do mesmo tipo têm aspectos diferentes, o que é um problema na classificação.

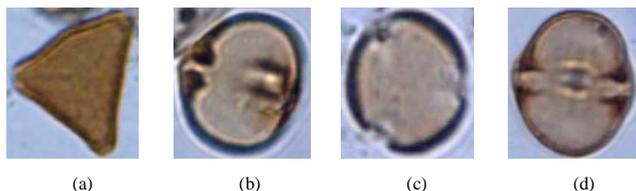


Figura 9. Imagens classificadas como *Protium*.

A Figura 10 mostra imagens que foram classificadas como *Schinus*. As imagens (b) e (c) estão classificadas corretamente. As imagens (a) e (d) são do tipo *Protium*. Forma, textura e cor são características muito semelhantes entre esses dois tipos, fazendo com que a classificação seja mais difícil.

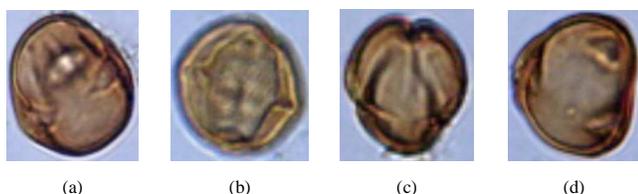


Figura 10. Imagens classificadas como *Schinus*.

A Figura 11 mostra os resultados da classificação para o tipo *Serjania*. As imagens (b) e (c) estão classificadas corretamente. A imagem (a) é do tipo *Poaceae* e a imagem (d) é do tipo *Schinus*. O erro na classificação pode-se observar que é em relação à forma como foi feita a captura das imagens. Por exemplo, as imagens (b) e (c) são do mesmo tipo, contudo existe uma variação na coloração, permitindo que o tipo *Serjania* apresente cores semelhantes as duas espécies *Poaceae* e *Schinus*.

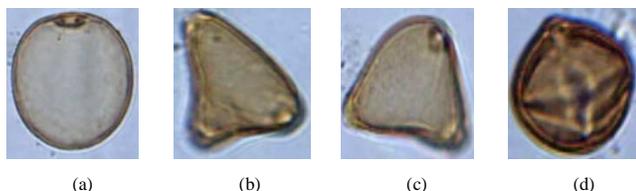


Figura 11. Imagens classificadas como *Serjania*.

Os resultados do tipo *Anadenanthera* foram todos classificados corretamente, em contrapartida, todas as imagens do tipo *Syagrus* foram classificadas erroneamente, sendo confundidas com imagens de outras espécies.

Com o algoritmo SMO 75% das imagens utilizadas no experimento foram classificadas corretamente. Este percentual já é um grande avanço na classificação polínica, pois [7] utilizando apenas dois gêneros conseguiu 97%, e como eles afirmam, quando aumenta o número de classe não é esperado que esta porcentagem de acerto se mantenha.

VI CONCLUSÃO

Este trabalho foi desenvolvido com o intuito de automatizar a classificação dos tipos polínicos. Foram realizados quatro experimentos utilizando métricas diferentes e algoritmos de aprendizagem diferentes. Com a análise do

desempenho dos algoritmos nas métricas impostas foi possível determinar os algoritmos que mais se destacaram positivamente. Em relação ao desempenho em porcentagem de acerto, o SVM obteve melhor resultado. Em relação à métrica Área sob a curva ROC e Média Ponderada da média F o algoritmo SVM obteve melhor resultado. Para o problema da classificação de grãos de pólen os algoritmos SVM e KNN são os mais indicados para o problema. Em relação ao tamanho de dicionário mais adequado a esse conjunto de dados, foi possível observar que o dicionário 2048 se mostrou mais adequado, pois apresentou bons resultados na maioria das métricas para diferentes classificadores. Para trabalhos futuros espera-se que sejam aplicadas novas formas de extração de atributos para as mais variadas classes de grãos de pólen e que sejam testados novos algoritmos de aprendizagem utilizando outras métricas.

REFERÊNCIAS

- [1] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." *Machine learning* 6.1 (1991): 37-66.
- [2] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." *Computer Vision-ECCV 2006*. Springer Berlin Heidelberg, 2006. 404-417.
- [3] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern recognition* 30.7 (1997): 1145-1159.
- [4] Costa, Clayton M., and Suann Yang. "Counting pollen grains using readily available, free image processing and analysis software." *Annals of botany* 104.5 (2009): 1005-1010.
- [5] Cruz-Roa, Angel, Juan C. Caicedo, and Fabio A. González. "Visual pattern mining in histology image collections using bag of features." *Artificial Intelligence in Medicine* 52.2 (2011): 91-106.
- [6] Csurka, Gabriella, et al. "Visual categorization with bags of keypoints." *Workshop on statistical learning in computer vision, ECCV*. Vol. 1. 2004.n
- [7] Damian, M., et al. "Pollen classification of three types of plants of the family Urticaceae." *Proc. of the 12th Portuguese Conference on Pattern Recognition, Aveiro, Portugal*. 2002.
- [8] France, I., et al. "A comparative study of approaches to automatic pollen identification." *Proceedings of British Machine Vision Conference*. 1997.
- [9] Hripcsak, George, and Adam S. Rothschild. "Agreement, the f-measure, and reliability in information retrieval." *Journal of the American Medical Informatics Association* 12.3 (2005): 296-298.
- [10] Kim, Kyungnam, et al. "Real-time foreground-background segmentation using codebook model." *Real-time imaging* 11.3 (2005): 172-185.
- [11] Louveaux, J., Maurizio. A., Vorrwohl, G. (1970) Methods of melissopalynology, *Bee World* 51, 125-131.
- [12] Paes, A. T. "Por dentro da estatística." *Einstein: Educ. Contin. Saúde* 6 (2008): 107-108.
- [13] Quinlan, John Ross. *C4. 5: programs for machine learning*. Vol. 1. Morgan kaufmann, 1993.
- [14] Schapire, Robert E. "The boosting approach to machine learning: An overview." *LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-* (2003): 149-172.
- [15] Situ, Ning, et al. "Malignant melanoma detection by bag-of-features classification." *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008.
- [16] Vapnik, Vladimir. *The nature of statistical learning theory*. Springer, 1999.
- [17] Tamaki, Toru, et al. "A system for colorectal tumor classification in magnifying endoscopic NBI images." *Computer Vision-ACCV 2010*. Springer Berlin Heidelberg, 2011. 452-463.
- [18] Wolff, L.F.; Reis, V.D.A.; Santos, R.S.S. Abelhas melíferas: bioindicadores de qualidade ambiental e de sustentabilidade da agricultura familiar de base ecológica. Pelotas: Embrapa Clima Temperado, 38p., 2008.