Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

Relatório Final de Atividades

Caracterização de Bioimagens

vinculado ao projeto

Métodos e técnicas para exploração e análise de bioimagens

Thiago Pereira Colonhezi
Voluntário
Análise e Desenvolvimento de Sistemas
Data de ingresso no programa: 08/2011
Orientador: Prof. Dr. Fabrício Martins Lopes
Co-orientador: Prof. Me. Pedro H. Bugatti

Área do Conhecimento: 1.03.04.00-2 Sistemas de Computação

CAMPUS CORNÉLIO PROCÓPIO 2012

THIAGO PEREIRA COLONHEZI

CARACTERIZAÇÃO DE BIOIMAGENS

Relatório Pesquisa do Programa de Iniciação Tecnológica da Universidade Tecnológica Federal do Paraná.

CORNÉLIO PROCÓPIO, 2012

Sumário

Introdução

Nos últimos anos ocorreram avanços tecnológicos nas mais diferentes áreas do conhecimento, entre elas a aquisição de imagens e vídeos ganhou destaque e atualmente é possível ter acesso a equipamentos razoavelmente sofisticados e com boa resolução de imagens e vídeos a preços populares.

Como consequência natural dessa popularização dos equipamentos de aquisição de imagens e vídeos, houve uma explosão na captação e armazenamento de conteúdo multimídia (imagens e vídeos).

Logo, indexar e recuperar imagens e vídeos em meio a um conjunto enorme de arquivos passou a ganhar destaque na comunidade acadêmica/científica e também em empresas privadas, como por exemplo, o Google. Em outras palavras, existe atualmente a necessidade explícita de indexar e recuperar imagens e vídeos em meio a um volume massivo de dados.

Claramente, essa recuperação pode se dar de forma textual, por meio da inclusão de identificadores conhecidos como *"tags",* a qual apresenta a vantagem de se poder indexar e recuperar rapidamente os conteúdos multimídia. No entanto, essa indexação exige que uma informação textual seja incluída ao conteúdo multimídia. Caso essa informação não seja incluída, o conteúdo não será indexado e, por consequência, não será recuperado.

Logo, para se adotar essa estratégia de indexação é necessário que o domínio (informações sobre a imagem) seja corretamente estabelecido, ou seja, a atribuição das *tags*. Esse tema está diretamente relacionado com um dos cinco grandes desafios identificados pela Sociedade Brasileira de Computação (SBC): gestão da informação em grandes volumes de dados multimídia distribuídos [1].

Uma forma que independe dessa limitação é a recuperação de conteúdo multimídia baseado em seu conteúdo e não necessita de atribuições de rótulos que a identifiquem. A imagem/vídeo é recuperada de acordo com a pergunta realizada a base de dados.

Essa nova forma de recuperação de conteúdo multimídia normalmente é baseada na extração de características da fonte, as quais são utilizadas para estabelecer os critérios de recuperação.

No Brasil e em outras partes do mundo, temos a grande necessidade em se conhecer a flora de forma rápida e ágil. A flora brasileira é considerada uma das mais ricas, com mais de 56.000 espécies, representando cerca de 19% da flora mundial [2]. O Brasil também é o país com a maior diversidade biológica do planeta, com alto índice de espécies endêmicas. Essa diversidade biológica é muito expressiva tanto em relação às potencialidades genéticas como em relação ao número de espécies e de ecossistemas. Considerando a biodiversidade vegetal, a Floresta Amazônica é detentora da maior reserva de plantas medicinais do mundo.

O conhecimento exato de uma espécie de planta, distribuição geográfica ou a utilização, é essencial para o desenvolvimento da agricultura. Muitas das vezes essas informações não estão disponíveis de forma prática, pois é necessário recorrer a grandes catálogos de espécies limitando o conhecimento aos interessados profissionais, professores e cientistas. As plantas constituem uma dificuldade até mesmo para os agricultores e cientistas, devido à variabilidade de sua forma, muito maior do que a biologia animal. Por isso, surge a necessidade de métodos computacionais que auxiliem na identificação das espécies utilizando uma imagem de recuperação.

Há atualmente, diversas empresas e cientistas que buscam uma forma de realizar buscas por imagem, visando analisar o seu conteúdo e características e através disso retornar um resultado esperado. Com uma planta não seria diferente, seria extremamente útil para um pesquisador, utilizando um celular, obter uma foto de uma planta e realizar buscas em um banco especifico e obter resposta útil em tempo ágil, não necessitando mais de busca manual em grandes catálogos. Para que isso seja possível, é necessário analisar a imagem e retirar características de objetos ou regiões, i.e. sobre geometria, textura e cor.

Este trabalho aborda o tema de extração de características com ênfase na textura utilizando os descritores de Haralick [3] e posteriormente realiza uma classificação com o objetivo de conhecer a classe de uma determinada folha de exemplo.

Para isso, foi utilizado um banco de dados disponibilizado pela organização francesa ImageClef [4] <*http://www.imageclef.org/*>, o qual promove anualmente um evento o qual lança desafios à comunidade científica internacional relacionado ao processamento de imagens. Na ultima edição, 2011, disponibilizaram uma nova modalidade de reconhecimento de espécies de árvores da Europa.

Também objetivo deste trabalho utilizar o banco de dados ImageClef para obter as características das folhas e aplicar diferentes classificadores disponíveis no software de mineração de dados WEKA [5] para o estudo e análise dos melhores classificadores para o problema em questão.

Materiais e Métodos

Conceitos básicos sobre Imagens

Antes de realizar algumas etapas deste projeto, foi necessário um estudo específico de alguns conceitos sobre imagens digitais.

O Termo *imagem monocromática*, ou simplesmente *imagem*, refere-se à função bidimensional de intensidade da luz $f(x,y)$ onde x e y denotam as coordenadas espaciais e o valor de f em qualquer ponto é proporcional ao brilho da imagem no ponto [6].

Portanto uma imagem digital pode ser considerada uma matriz, na qual os elementos mostram um ponto na imagem que corresponde aos níveis de intensidade de luz da imagem. Os elementos desta matriz são chamados de *elementos da imagem* ou *pixels*.

Vizinhança

O relacionamento básico entre os pixels pode ocorrer ser de três formas distintas: Conectividade 4 – onde são verificados os quatros vizinhos de um determinado pixel central; Conectividade de 8 – onde são verificado os 8 vizinhos de um pixel central e a Conectividade de m – onde é modificado a conectividade de 8 para conexões onde existem diversos caminhos diferentes e é necessário contornar estes obstáculos para se identificar a borda.

Com a conectividade é possível determinar rótulos em objetos ou regiões da imagem, para isso, é necessário percorrer toda a imagem pixel a pixel identificando características em comum, como intensidade dos canais ou dos níveis de cinza. Após definidos os diferentes rótulos é possível calcular distancias ou enumerar a quantidade de itens que a imagem possui, dentre outras utilidades. Neste trabalho foi utilizada a conectividade para calcular a área das folhas, uma das características utilizadas.

Histograma

Outro conceito utilizado foi o histograma o qual representa o agrupamento de pixels que possuem a mesma de intensidade. O histograma caracteriza uma imagem. Contudo contem muitos valores e com isso pode conter redundâncias [7], por isso utilizamos os seguintes valores: media mediana e desvio padrão.

Textura

Além do histograma, foi utilizado a textura que normalmente é independente de posição, orientação, tamanho, forma e brilho da imagem. A definição de textura pode ser dita como: a disposição ou característica dos elementos constituintes de algo, especialmente no que se refere à aparência superficial ou à qualidade táctil. Para imagens a textura é uma característica representativa da distribuição espacial dos elementos ou pixels de uma imagem em uma região. Portanto obtermos características a partir de textura é um valor que representa a variação dos níveis de cinza de uma região ou imagem.

Extração de características

Chama-se extração de características todo o conjunto de operações de processamento e análise de imagens realizadas com a finalidade de obter valores numéricos que caracterizam as imagens ou partes delas. Também pode ser definido como sendo a captura das informações mais relevantes de um dado fornecido como entrada [8].

Essas características podem ser agrupadas em quatro grandes categorias: Características morfológicas, Cromáticas, Texturas e Estruturais ou Contextuais. Dentre estas adotamos as características morfológica, cromáticas e de textura.

As características morfológicas são as medidas que compõe a imagem, não levando em conta a intensidade dos pixels podendo ser calculadas através de imagens binárias geradas de uma imagem colorida, como exemplo: circularidade, largura, perímetro e área, dentre estas a área foi selecionada neste trabalho.

Características cromáticas são as que descrevem a cor, ou composição espectral da radiação emitida ou refletida pelos objetos, quantificada pela intensidade dos pixels em diferentes bandas espectrais ou cores. Neste trabalho, como descrito anteriormente, foram utilizadas as medidas estatísticas obtidas através do histograma da imagem. Estas medidas foram: mediana, média e desvio padrão. Já com relação à textura, foram utilizadas características obtidas por meio da aplicação de matrizes de co-ocorrência e dos descritores de Haralick.

Matriz de co-ocorrência

Uma forma rápida para demonstrar o uso da textura são os descritores de Haralick. Em 1973 Haralick apresentou a matriz de co-ocorrência e descritores de texturas para classificação de 6 tipos de rochas [3].

A matriz de co-ocorrência (Figura 1) de textura considera a relação entre dois pixels por vez, um chamado de pixel referência e o outro de pixel vizinho. O pixel vizinho escolhido em 4 diferentes ângulos 0º, 45º, 90º e 135º (direita, esquerda acima, abaixo ou diagonal) assim gerando 4 matrizes de co-ocorrência para análise da textura. Há também a distancia em pixel entre o pixel de referência e o vizinho.

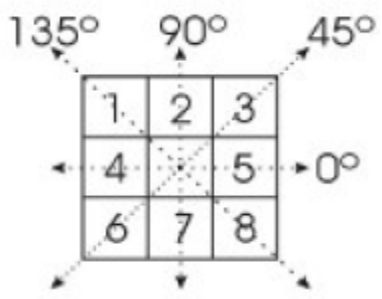

Figura 1- Vizinhança de um pixel central e os diferentes ângulos possíveis. **[3]**

A matriz de co-ocorrência pode ser especificada por uma matriz de frequências relativas na qual dois elementos de textura vizinhos, separados por uma distancia *d* em uma orientação Θ ocorrem em uma imagem, um com uma intensidade 1 e outra com uma intensidade j. Desta forma, a matriz representa em cada elemento o número de vezes que ocorreu uma transição do nível de cinza de A para B considerando a distancia e a direção.

Descritores de Haralick

A partir da geração das matrizes de co-ocorrência, calculam-se os 13 descritores propostos por Haralick, os quais estão relacionado abaixo de 1 a 13:

1. Segundo momento angular (*Angular Second Moment*)

$$f1 = \sum_i \sum_j p(i,j)^2 \tag{1}$$

2. Contraste (*Contrast*)

$$f2 = \sum_{n=0}^{Ng-1} n^2 \left\{ \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j)_{|i-j|=N} \right\} \tag{2}$$

3. Correlação ou Variância (*Correlation*)

$$f3 = \frac{\sum_i \sum_j (ij).p(ij) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3}$$

Onde μ_x, μ_y, σ_x e σ_y são as medias e desvio padrão de p_x e p_y.

4. Variância da soma (*Sums of Square: Variance*)

$$f4 = \sum_i \sum_j (i-\mu)^2 p(i,j) \tag{4}$$

5. Momento Diferança Inverso (*Inverse Difference Moment*)

$$f5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \tag{5}$$

6. Soma das Médias (*Sum average*)

$$f6 = \sum_{i=2}^{2Ng} i p_{x+y}(i) \tag{6}$$

7. Soma da Variação (*Sum variance*)

$$f7 = \sum_{i=2}^{2Ng} (i - f8)^2 p_{x+y}(i) \tag{7}$$

8. Soma da Entropia (*Sum entropy*)

$$f8 = \sum_{i=2}^{2Ng} p_{x+y}(i) \log\{p_{x+y}(i)\} \tag{8}$$

9. Entropia (*Entropy*)

$$f9 = -\sum_{i}\sum_{j} p(i,j) \log\big(p(ij)\big) \tag{9}$$

10. Diferença da Variação (*Difference variance*)

$$f10 = variance\ of\ p_{x-y} \tag{10}$$

11. Diferença da Entropia (*Difference entropy*)

$$f11 = \sum_{i=0}^{Ng-1} p_{x-i}(i) \log\{p_{x-i}(i)\} \tag{11}$$

12. Informação da Medida Correlação 1 (*Information measure of correlation 1*)

$$f12 = \frac{HXY - HXY1}{\max\{HX, HY\}} \tag{12}$$

13. Informação da Medida Correlação 2 (*Information measure of correlation 2*)

$$f13 = (1 - \exp[-2.0(HXY2 - HXY)])^{\frac{1}{2}} \tag{13}$$

$$HXY = -\sum_{i}\sum_{j} p(i,j) \log(p(i,j))$$

Após obter as características são necessários algoritmos específicos para a classificação. Para tal processo foi conduzido utilizando o software WEKA [5] que é um produto da Universidade de Waikato (Nova Zelândia) e foi implementado pela primeira vez em sua forma moderna em 1997. O software foi desenvolvido utilizando a linguagem JAVA e contém uma interface gráfica para interagir com arquivos de dados e produzir alguns resultados visuais, tabelas e gráficos. Nele, foram utilizados algoritmos de classificação com relação às respectivas espécies de plantas disponíveis na base de imagens.

Classificação

A classificação é um processo de reconhecimento de padrões que cria um guia passo a passo com a maneira para determinar a saída de uma nova instância de dados. A árvore, ou processo, cria nós que representam um ponto de decisão que pode ser tomado de acordo com o dado de entrada e conforme se move é possível determinar uma saída prevista.

Tipicamente em um processo de aprendizagem supervisionada em reconhecimento de padrões, após o pré-processamento e a formatação, os dados são fragmentados em dois subconjuntos, denominados base de treinamento e base de testes.

Numa primeira etapa o algoritmo de conhecimento é aplicado à base de treinamento. Com isso se obtém um modelo, que de certa forma representa o conhecimento e as regras inferidas. Numa segunda etapa o modelo obtido é aplicado ao fragmento da base de dados denominado base de testes. Como base de testes também é precisamente rotulada é possível medir a taxa de acerto do modelo, comparando-se o resultado obtido com a rotulação disponível na base de testes.

A técnica empregada neste trabalho para a avaliação do processo de classificação foi a de validação cruzada [9] onde consiste em dividir a base de dados em N partes chamadas também de *folds*. Destas n-1 partes são utilizadas para o treinamento e uma serve como base de testes. O processo é repetido N vezes, de forma que cada parte seja uma vez usada como conjunto de testes. Ao final, a correção total é calculada pela média dos resultados obtidos em cada etapa, obtendo-se assim uma estimativa da qualidade do modelo de conhecimento gerado e permitindo análises estatísticas.

Neste trabalho foram utilizados os seguintes classificadores: Naive Bayes, Multi class Classifier, Multilayer Perceptron, Adaboost com Naive Bayes, Adaboost com J48 e Adaboost com Multilayer.

- Naive Bayes: a decisão Bayesiana é fundamental para a introdução em problemas de classificação de padrões [10]. É a forma mais simples da rede bayesiana, onde todos os atributos são condicionalmente independentes, ou seja, a formação do evento não é uma informativa sobre nenhum outro.

- AdaBoost é utilizado em conjunto com outro classificador, o que pode reduzir significantemente o nível de erros. O método trabalha de certa forma aumentando várias vezes a distribuição dos dados de treinamento, fazendo com que o algoritmo de classificação tenha um conjunto de treinamento mais completo auxiliando a criação do modelo.

- Multilayer Perceptron: tem sido aplicada com sucesso em uma variedade de áreas, desempenhando tarefas como: classificação de padrões, controle e processamento de sinais. O Multilayer Perceptron é baseado em redes neurais. As redes neurais é um modelo que simula o cérebro em realizar determinadas tarefas. É um processamento distribuído que adquire conhecimento a respeito dos dados de entrada, ao decorrer do processo. São constituídos por um conjunto de nós fonte, os quais formam a camada de entrada da rede e os nós de saída. O multilayer Perceptron, para diminuir as limitações do conhecimento, utiliza três recursos diferenciais: O modelo de cada neurônio (nós) inclui uma função de ativação não linear o que o diferencia. A rede possui uma ou mais camadas escondidas entre a camada de entrada e saída. O terceiro diferencial, a rede possui um alto grau de conectividade entre as camadas [11].

Descrição da Base de Imagens

O presente trabalho teve como entrada os dados do banco de dados do ImageClef [12] edição de 2011, no qual foram disponibilizada pela primeira vez um banco deste gênero descrito como identificação de plantas (*Plant identification*).

O banco de dados possui folhas de 71 espécies de árvores da área mediterrânea francesa. Contém aproximadamente 5436 imagens subdivididas em 3 diferentes tipos de imagens: digitalização, fotos que simulam a digitalização e fotos naturais.

Digitalização: esta classe de imagens possui 3070 imagens, as quais foram coletadas entre julho e setembro de 2009 e junho a outubro de 2010. Foi criado por pesquisadores, biólogos e voluntários pertencentes a uma rede social chamado de *Telabotanica <http://www.tela-botanica.org/site:accueil?langue=en>*. A Figura 2 exibe 2 exemplos de folhas digitalizadas.

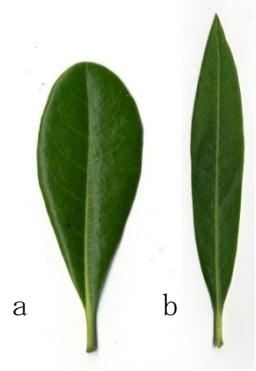

Figura 2 - Imagem de duas folhas digitalizadas. a) Pittosporum tobira e b) Nerium oleander.

Fotos como digitalização: são disponibilizadas 897 imagens para a segunda classe, as quais são similares a uma digitalização e podem apresentar o fundo um pouco uniforme e apresentar sombras nas folhas. A Figura 3 exibe dois exemplos de imagens desta classe.

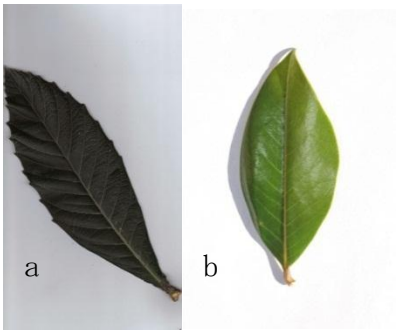

Figura 3 - Imagem de duas folhas imagem que simulam a digitalização. a) Eriobotrya japonica e a b) Magnolia grandiflora.

Fotos naturais: a terceira classe de imagens apresenta folhas em paisagens naturais e existem 1469 imagens, tiradas diretamente nas árvores. As imagens podem apresentar como fundo: troncos, várias folhas, a terra ou o céu. A Figura 4 exibe dois exemplos dessa classe.

Figura 4 - Imagem de duas folhas naturais. a) Ilex aquifolium a b) Vitex agnus-castus.

Ferramenta WEKA

A ferramenta Weka [5] é um software do tipo *open source* para a mineração e classificação de dados. Um classificador (ou modelo de classificação) é utilizado para identificar a classe à qual pertence uma determinada observação de uma base de dados, a partir de suas características (seus atributos).

O Weka trabalha com arquivos de entrada do tipo ARFF, que corresponde a um arquivo de texto contendo um conjunto de observações, precedido por um pequeno cabeçalho. O cabeçalho é utilizado para fornecer informações a respeito dos campos que compõem o conjunto de observações. Dessa forma, antes da mineração de dados, a ferramenta pode verificar alguma inconsistência na base de dados e sinalizá-la.

O arquivo possui um cabeçalho que é definido por @relation, posteriormente terá a sequencia de atributos definido por @attribute e para finalizar terá o comando @data para definir o início das observações. Por padrão, o ultimo atributo apresentado na relação será o atributo classe, porém isso pode ser alterado na interface do programa.

Validação dos Dados

Quando se desenvolve sistemas que envolvem a detecção, diagnósticos ou previsão de resultados é importante validar os seus resultados de forma a qualificar seu poder discriminativo e identificar um método como bom ou não para determinada análise.

Precisão e Sensitividade (Recall)

Ao utilizar precisão e recall [13], o conjunto de rótulos possíveis para uma determinada instância é dividido em dois subgrupos, um dos quais é considerado relevante para os objetivos da métrica. Recall é então calculado como a fração de instâncias corretas entre todas as instâncias que realmente pertencem ao subconjunto relevante.

A precisão é a fração de instâncias corretas entre aqueles que o algoritmo considera pertencer ao subconjunto relevante. A precisão pode ser vista como medida de exatidão ou fidelidade, enquanto o recall é uma medida de completude.

Curvas ROC

A curva ROC [14] (Figura 2) foi desenvolvida por engenheiros elétricos e engenheiros de sistemas de radar durante a Segunda Guerra Mundial para detectar objetos de inimigos em cambo de batalha. A análise ROC tem sido utilizada em medicina, radiologia, psicologia e outras áreas por muitas décadas e, mais recentemente, foram introduzidas a áreas como aprendizado de máquina e mineração de dados. A curva ROC é útil em domínios nos

quais existe uma grande desproporção entre as classes ou quando se deve levar em consideração diferentes custos/benefícios para os diferentes erros/acertos de classificação.

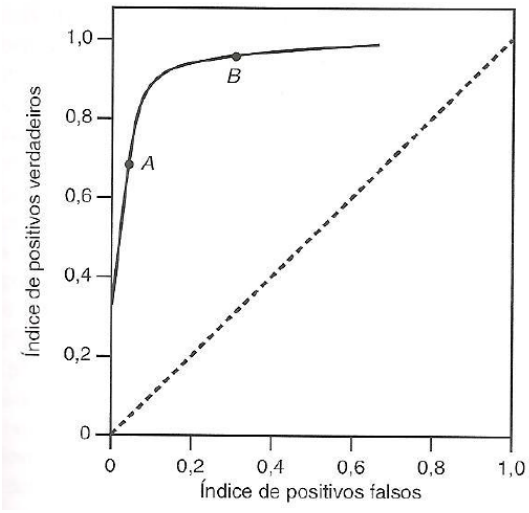

Figura 5 – Curva característica de operação do receptor (Curva ROC).

O Gráfico ROC é baseado na probabilidade de detecção, ou taxa de verdadeiros positivos e na probabilidade de falsos alarmes ou taxa de falsos positivos. Alguns pontos no gráfico merecem destaque. O ponto (0,0) representa a estratégia de nunca classificar um exemplo como positivo. Modelos que correspondem a esse ponto não apresentam nenhum falso positivo, mas também não conseguem classificar nenhum verdadeiro positivo. A estratégia inversa, de sempre classificar um novo exemplo como positivo, é representada pelo ponto (1,1). O ponto (0,1) representa o modelo perfeito, onde todos os exemplos positivos e negativos são corretamente classificados. O ponto (1,0) representa o modelo que sempre faz predições erradas.

Resultados e Discussões

Primeiramente foi implementado em JAVA [15] um software capaz de abrir a imagem e extrair as características descritas anteriormente.

O jImageFature, software o qual foi desenvolvido e se apresenta como um dos resultados deste trabalho, pode ser encontrado no Google Code, no seguinte link: <http://code.google.com/p/jimagefeature/>. Com este software é possível abrir uma imagem em questão e visualizar os descritores de Haralick, estatísticas do histograma, realizar medidas entre as características desejadas de várias imagens em uma pasta, bem como exportar estas características com o padrão ARFF, suportado pelo WEKA.

O software foi organizado nos seguintes pacotes: Controle (Control), Interface gráfica (GUI) e Útil. No pacote de Controle, ficaram as classes pertinentes aos cálculos e processos necessários para a geração das características e resultados. Estas classes são: Coocorrencia, responsável pela geração das matrizes de co-ocorrência; Haralick, responsável pelos cálculos dos descritores; Histogramas (histograms), responsável pela geração dos histogramas e cálculos do mesmo; Forma (shape), responsável pelo cálculo da área das folhas; Metricas (metrics) onde é possível obter medidas de distância dos vetores desejados; Características (features), classe responsável pela comunicação com as demais classes do software; Opções (options), onde são controladas as opções selecionadas pelo usuário ao

gerar os dados e o ultimo pacote, o de interface gráfica onde contém as classes responsável pela interação com o software.

O ultimo pacote, Útil, possui a classe responsável pela gravação do arquivo de texto gerado pelo software.

Para um melhor controle dos atributos gerados para a classificação, foi definido o vetor de características podendo na sua forma completa estar na seguinte ordem:

1. Contrast – 0º
2. Correlation – 0º
3. Sums of Square: Variance – 0º
4. Inverse Difference Moment – 0º
5. Sum average – 0º
6. Sum variance – 0º
7. Sum entropy – 0º
8. Entropy – 0º
9. Difference variance – 0º
10. Difference entropy – 0º
11. Information measure of correlation 1 – 0º
12. Information measure of correlation 2 – 0º
13. Angular Second Moment – 45º
14. Contrast – 45º
15. Correlation – 45º
16. Sums of Square: Variance – 45º
17. Inverse Difference Moment – 45º
18. Sum average – 45º
19. Sum variance – 45º
20. Sum entropy – 45º
21. Entropy – 45º
22. Difference variance – 45º
23. Difference entropy – 45º
24. Information measure of correlation 1 – 45º
25. Information measure of correlation 2 – 45º
26. Angular Second Moment – 90º
27. Contrast – 90º
28. Correlation – 90º
29. Sums of Square: Variance – 90º
30. Inverse Difference Moment – 90º
31. Sum average – 90º
32. Sum variance – 90º
33. Sum entropy – 90º
34. Entropy – 90º
35. Difference variance – 90º
36. Difference entropy – 90º
37. Information measure of correlation 1 – 90º
38. Information measure of correlation 2 – 90º
39. Angular Second Moment – 135º
40. Contrast – 135º
41. Correlation – 135º
42. Sums of Square: Variance – 135º
43. Inverse Difference Moment – 135º
44. Sum average – 135º
45. Sum variance – 135º
46. Sum entropy – 135º
47. Entropy – 135º
48. Difference variance – 135º
49. Difference entropy – 135º
50. Information measure of correlation 1 – 135º
51. Information measure of correlation 2 – 135º
52. Median
53. Pixels-Average
54. Standard-Deviation
55. Area

Com o vetor definido, foi necessário implementar a exportação dos dados para o software WEKA, conforme foi descrito anteriormente.

Conforme citado na Seção de Classificação, no presente trabalho foram utilizados os classificadores Multilayer Perceptron (Figura 9), Multi Class Classifier (Figura 8), Naive Bayes (Figura 11), Adaboost-J48 (Figura 7), Adaboost-Multilayer (Figura 12) e Adaboost Naive Bayes (Figura 10). Com os resultados obtidos foi possível observar as diferenças entre os classificadores a partir das análises das curvas ilustradas no Figura 6, no qual o eixo das abscissas são representadas as respectivas classes de imagens folheares e no eixo das ordenadas a média de acertos obtidos pelos classificadores.

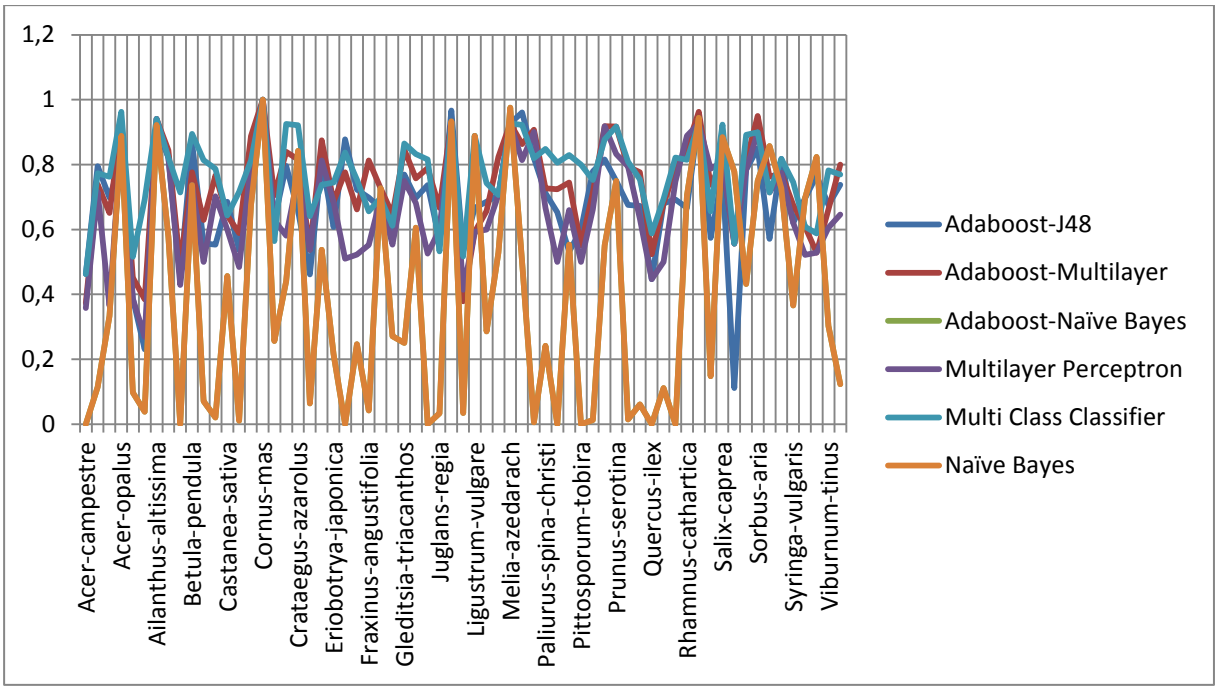

Figura 6 - Acertos obtidos por cada classificador nas classes disponíveis.

Retirando o classificador Naive bayes, todos os outros apresentaram a média de acerto superior a 70%.

A análise dos classificadores foi realizada por meio do percentual de acertos na identificação das classes das folhas. A Tabela 1 exibe os percentuais de acertos obtidos de cada classificador.

Para comparar os classificadores (Tabela 1) deve atentar para os valores de Verdadeiro Positivo, ROC, precisão e Sensibilidade que quanto mais próximo de 1 melhor foi a classificação e para os valores de Falso Positivo, quanto mais próximo de zero, melhor a classificação, mostrando que esta medida é complementar.

Tabela 1 – Percentuais de acertos obtidos pelos classificadores.

Classificador	**Verdadeiro Positivo**	**Falso Positivo**	**Precisão**	**Sensibilidade (recall)**	**ROC Área**
Naive Bayes	0,294	0,011	0,286	0.294	0,868
Multilayer Perceptron	0,649	0,007	0,656	0,649	0,918
Multi Class Classifier	0,765	0,005	0,769	0,765	0,963
Adaboost-J48	0,688	0,007	0,685	0,688	0,953
Adaboost Multilayer Perceptron	0,725	0,006	0,728	0,725	0,967
Adaboost Naive Bayes	0,294	0,011	0,286	0,294	0,868

O valor de verdadeiro positivo indica a proporção de casos verdadeiros entre todos os casos com teste positivo. Portanto, quanto mais próximo de 1 melhor será o classificador. Já para falso positivo, indica a proporção de casos falsos entre todos os casos com teste falso, portanto quanto menor, melhor será o classificador.

As próximas figuras exibem separadamente as curvas de acertos por classe para cada classificador:

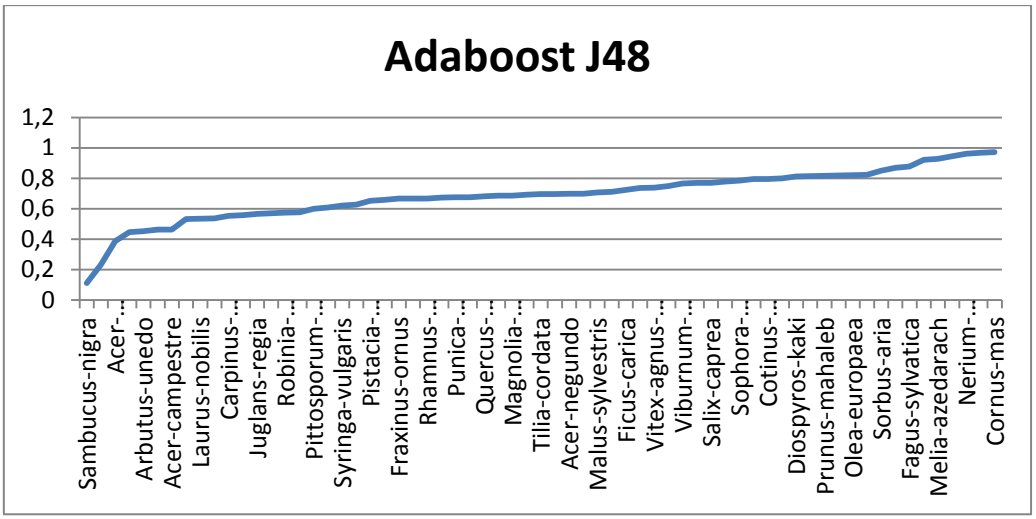

Figura 7 - Curvas de acerto para o classificador Adaboost J48

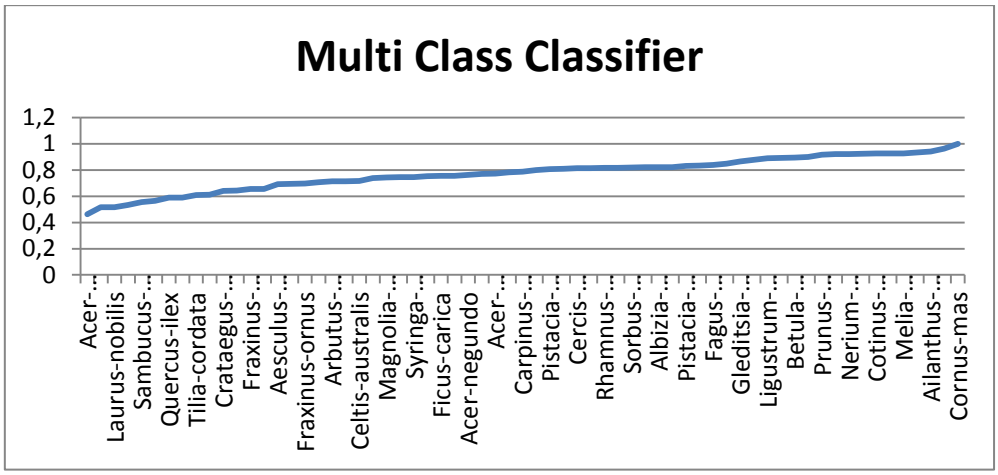

Figura 8 - Curvas de acerto para o classificador Multi Class Classifier

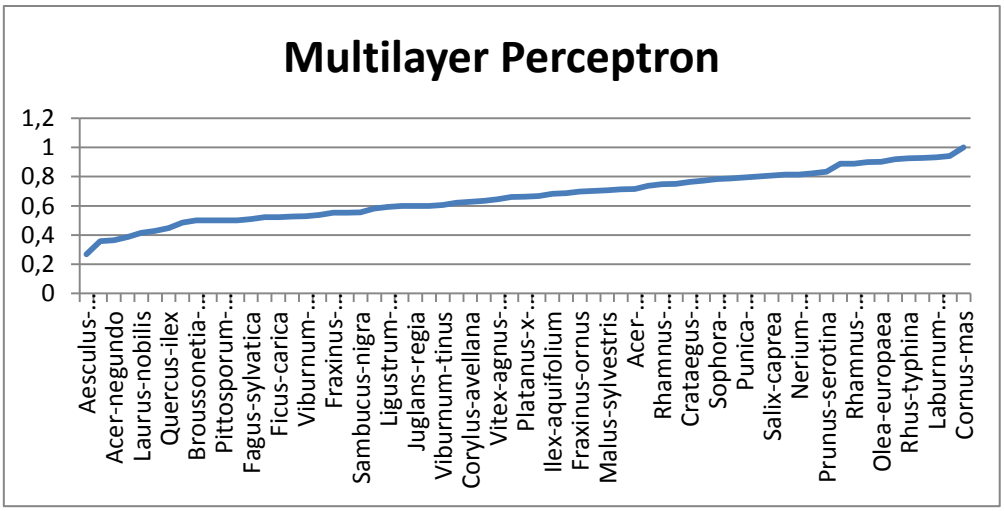

Figura 9 - Curvas de acerto para o classificador Multilayer Perceptron

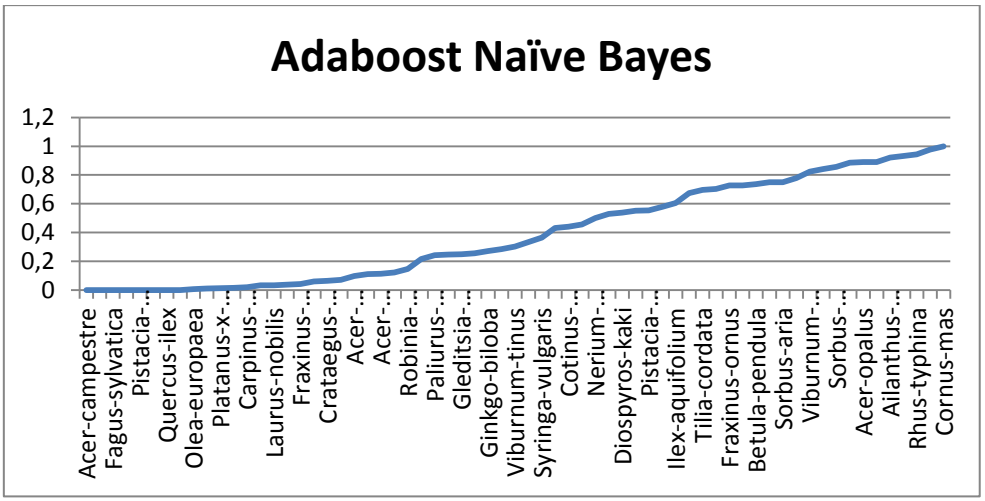

Figura 10 - Curvas de acerto para o classificador Adaboost Naive Bayes

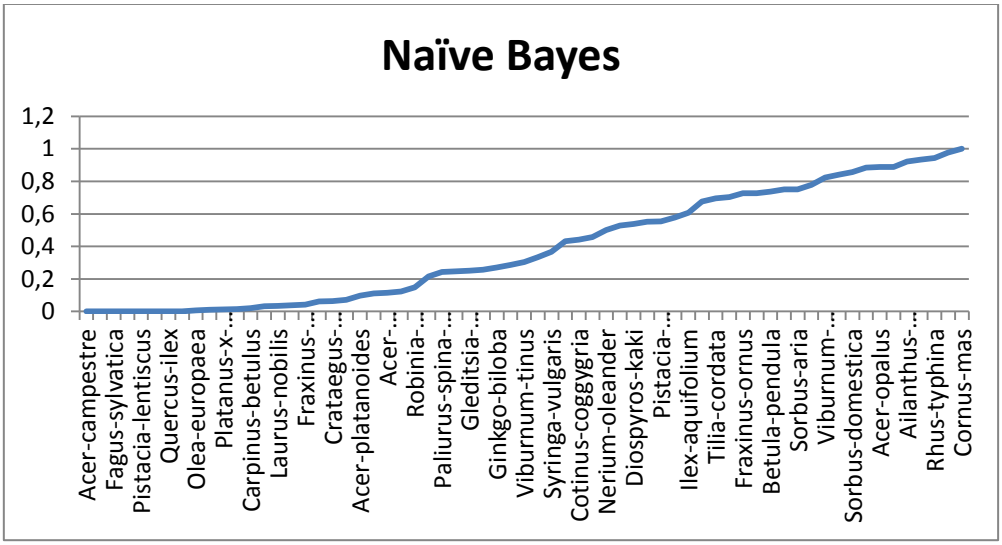

Figura 11 - Curvas de acerto para o classificador Naive Bayes

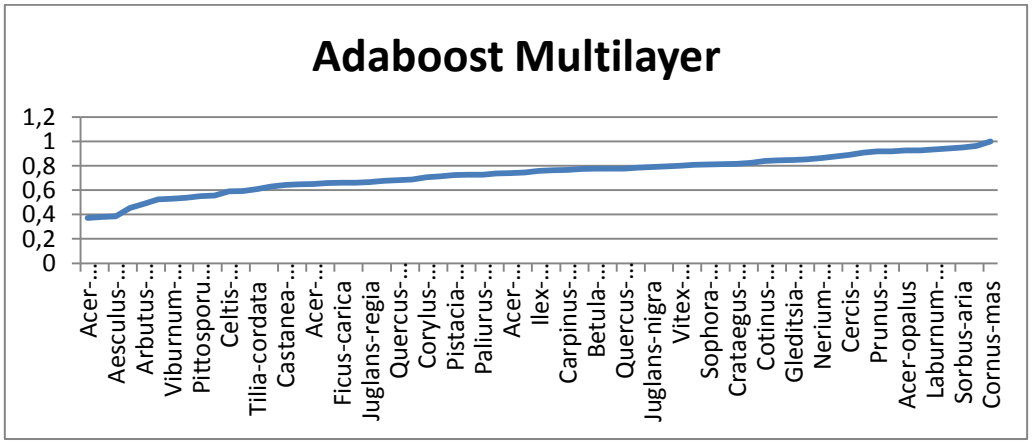

Figura 12 - Curvas de acerto para o classificador Adaboost Multilayer Perceptron

Os resultados obtidos indicam que os classificadores desempenham uma função crucial para o processo de identificação.

Dentre os classificadores analisados a média de acertos obtidos podem organizar na seguinte ordem de precisão: Multi Class Classifier – 77%, Adaboost Multilayer 73%, Adaboost J48 - 68% e Multilayer Perceptron - 66%. Os classificadores Naive Bayes e o Adaboost Naive Bayes obtiveram o mesmo resultado em 39%.

Em relação às classes, foi observado que a Cornus Mas (Figura 12 – A), com 35 instancias foi a classe mais simples para a classificação. O único classificador que não obteve 100% de acerto foi o AdaboostJ48, com 97,1%.

A classe considerada mais complexa, i.e com menor taxa de acerto entre os classificadores, foi a Aesculus Hippocastanum (Figura 12 – B), a qual obteve uma média de acertos em 27% em um total de 26 instancias. O classificador com a maior precisão para esta classe foi o Multi Class Classifier com 69% de acerto sendo que os demais não atingiram nem 50%. Em segundo lugar ficou o classificador Multilayer Perceptron com 38,5% de acertos, é importante destacar que o algoritmo Adaboost apresentou uma melhoria de 43.1% (26.9% para 38.5%) nos acertos obtidos, nessa classe.

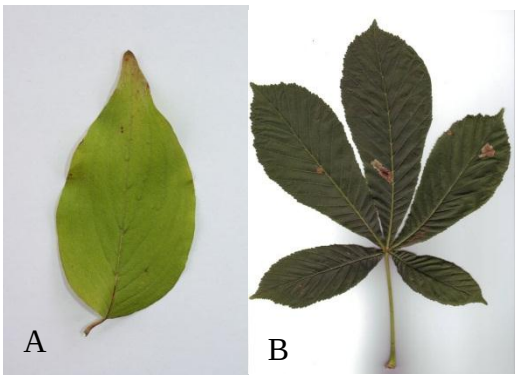

Figura 12 – Folha classe (a) Cornus Mas, (b) Aesculus Hippocastanum

Dentre as classes podemos visualizar que possuem diferentes níveis de dificuldades para a classificação a Tabela 2 apresenta as três classes mais fáceis, i.e. apresentaram maior índice de acerto. A Tabela 3 mostra as três classes mais complexas dentre as classes analisadas.

Tabela 2 – Classes com maiores índices de acertos durante a classificação

	Cornus-mas	**Melia-azedarach**	**Rhus-typhina**
Multi Class Classifier	1	0,927	0,926
Adaboost J48	0,971	0,927	0,944
Multilayer Perceptron	1	0,927	0,926
Adaboost Naïve Bayes	1	0,976	0,944
Naïve Bayes	1	0,976	0,944
Adaboost Multilayer Perceptron	1	0,927	0,963

Tabela 3 – Classes com menores índices de acertos durante a classificação

	Aesculus-hippocastanum	**Acer-campestre**	**Laurus-nobilis**
Multi Class Classifier	0,692	0,463	0,517
Adaboost J48	0,231	0,463	0,534
Multilayer Perceptron	0,269	0,358	0,414

Adaboost Naïve Bayes	0,038	0	0,034
Naïve Bayes	0,038	0	0,034
Adaboost Multilayer Perceptron	0,385	0,373	0,379

As dificuldades em se obter um padrão na classificação podem estar relacionadas às seguintes questões:

- Variação da forma de margem da folha: uma mesma classe pode apresentar diferentes margens dependendo da condição o qual foi obtido a imagem, um exemplo disto é a classe Quercus Ilex o qual é mostrado na Figura 13 - a;
- Numero de lóbulos: a mesma classe poderá apresentar diferentes números de lóbulos. Para a classe exemplificada abaixo na Figura 14 - b, Ficus Carica, poderá apresentar folhas com 3, 5 ou 7 lóbulos;

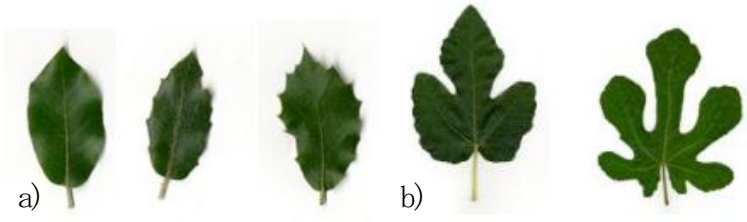

Figura 13 – Folhas da classe a) Quercus Ilex e b) Ficus Carica

- Forma geral e as variações de espessura: a forma poderá ser diferente entre as folhas, o mesmo acontece com a espessura. Na Figura 15 - a temos como exemplo folhas da classe Corylus Avellana;
- Coloração: A coloração poderá ser diferente, apresentando diferente níveis de cinza para uma mesma classe. Temos na Figura 15 - b a classe Cotinus Coggygria;

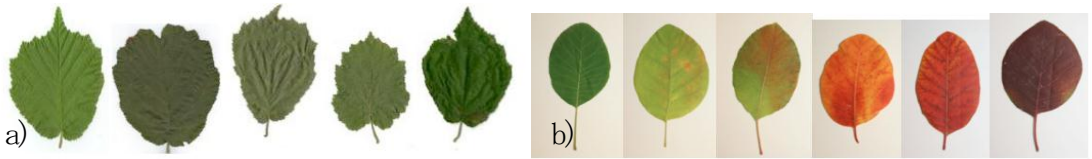

Figura 14 – Folhas da classe a) Corylys Avellana e b) Cotinus Coggygria

- Número de folhetos de uma folha composta: um exemplo de uma classe seria a Flaxinus Angustifolia mostrada na Figura 15 - a, o qual pode ter folhas de 3, 5, 7, 9 ou 11 folhetos;
- Reflexão da luz (utilização de flash): para as imagens do tipo como digitalização teremos os problemas em relação ao flash que poderá estar habilitado ou não, alterando assim a intensidade da luz que incide sobre a folha (Figura 15 - b).

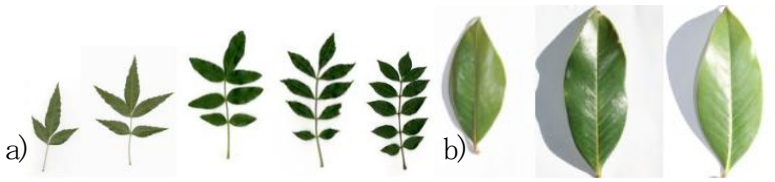

Figura 15 – Folhas da classe a) Flaxinus Angustifolia e b) Magnolia Grandiflora

Para as classes relacionadas acima, foram exibidos os resultados obtidos por cada classificador na Tabela 4. Cada classificador se comportou diferente em relação às dificuldades, destacando os classificadores Multi Class Classifier com a média de acerto em 70,5% e o Adaboost Multilayer Perceptron com a média de 70%. Novamente os classificadores Naive Bayes e Adaboost Naive Bayes apresentaram os piores resultados, apresentando uma taxa de acerto de 21%.

Tabela 4 – Classes e seus índices de acertos durante a classificação.

	Quercus ilex	**Ficus carica**	**Corylus avellana**	**Cotinus coggygria**	**Fraxinus angustifolia**	**Magnolia grandiflora**
Multi Class Classifier	0,588	0,754	0,564	0,925	0,656	0,743
Adaboost J48	0,447	0,723	0,628	0,796	0,698	0,686
Multilayer Perceptron	0,447	0,523	0,628	0,581	0,552	0,6
Adaboost Naïve Bayes	0	0,246	0,256	0,441	0,042	0,286
Naïve Bayes	0	0,246	0,256	0,441	0,042	0,286
Adaboost Multilayer Perceptron	0,524	0,662	0,705	0,839	0,813	0,657

Conclusões

Este trabalho teve como objetivo o desenvolvimento de um software o qual tivesse como saída dados para a análise de diferentes extratores de características, em conjunto com diversos classificadores encontrados na literatura, visando o reconhecimento de diferentes espécies folhares. Os resultados obtidos através de diversos testes mostram que as características selecionadas no trabalho podem ser utilizadas para realizar a anotação das classes para posterior indexação, pois em diversos classificadores foi possível obter resultados satisfatórios, evidenciando que tais representam as características de cada classe em questão. Para estas características foi possível selecionar dois (2) classificadores com melhor desempenho, i.e. apresentaram um maior número de acerto, foram: : Multi Class Classifier, Adaboost Multilayer Perceptron, ambos atingiram média de acertos acima de 70%. Já os classificadores Naive Bayes não alcançaram resultados satisfatórios, não atingindo nem 50% de precisão. Os resultados indicam que uma boa anotação das classes podem não atingir resultados satisfatórios quando é empregado o algoritmo de classificação incorreto.

Outro resultado deste trabalho foi o desenvolvimento de um software que implementa a extração das características mencionadas que pode ser livremente acessado através do Site: <http://code.google.com/p/jimagefeature/>.

Para trabalhos futuros poderá acrescentar algumas etapas para que seja possível eliminar ruídos e partes não importantes para a classificação, aumentando o foco nas características mais importantes de cada classe.

Desta forma, conclui-se que o trabalho desenvolvido pode ser utilizado em futuras pesquisas de classificação e busca de espécies folheares, adicionando uma contribuição na área que está em expansão. Com mais estudos e pesquisa, poderá tratar os ruídos nas características melhorando ainda mais a taxa de acertos.

Referências

1. SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. Grandes desafios 2009. **Sociedade Brasileira de Computação**, 2009. Disponivel em: <http://www.sbc.org.br/index.php?option=com_jdownloads&itemid=195&task=view.download&catid=50&cid=237>. Acesso em: 31 ago. 2012.
2. ANA MARIA GIULIETTI, R. M. H. L. P. D. Q. M. D. G. L. W. C. V. D. B. Biodiversidade e conservação das plantas no Brasil. **Megadiversidade**, p. 52-61, 2005.
3. SHANMUGAM, R. M. H. A. K. "Computer Classification of Reservoir Sandstones". **IEEE Transactions on Geoscience Electronics**, p. 171-177, 1973.
4. IMAGE Clef. **Image Clef - Image Retrieval in CLEF**, 2003-2011. Disponivel em: <http://www.imageclef.org/>. Acesso em: 11 jul. 2012.
5. WAIKATO, U. O. Weka Data Mining Software in Java. **Weka - The University of Waikato**. Disponivel em: <http://www.cs.waikato.ac.nz/ml/weka/>.
6. R.GONZALEZ, R. W. **Processamento de Imagens Digitais**. [S.l.]: [s.n.], 2004.
7. LOPES, F. M. UM MODELO PERCEPTIVO DE LIMIARIZAÇÃO DE IMAGENS DIGITAIS, 2003.
8. PIERRE A. DEVIJVER, J. K. Pattern recognition: a statistical approach. **Prentice/Hall International**, p. 448, 1982.
9. TESTE e validação (Mineração de dados). Disponivel em: <http://msdn.microsoft.com/pt-br/library/ms174493.aspx>.
10. RICHARD O. DUDA, P. E. H. E. D. G. S. **Pattern Classification**. Danvers: Wiley-Interscience, 2001.
11. HAYKIN, S. **Neural Networks and Learning Machines**. [S.l.]: Pearson, 1999.
12. FORUM, C. I. L. O. T. E. Image Clef Retrivial in CLEF. **Image Clef Retrivial in CLEF**. Disponivel em: <http://www.imageclef.org/2012>.
13. ALMEIDA, S. S. D. Uma Implementação de um Sistema de Contagem de Pessoas Baseado em Vídeo, 18 Dezembro 2010. Disponivel em: <http://www.decom.ufop.br/menotti/rp102/TrabalhoParcial-papers/01-PeopleCounting.pdf>. Acesso em: 01 set. 2012.
14. MARTINEZ, E. Z. A curva ROC para testes diagnósticos, Rio de Janeiro, p. 7-31, 2011.
15. ORACLE JAVA. **Oracle JAVA**. Disponivel em: <www.java.com/pt_BR/>.