UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS CORNÉLIO PROCÓPIO
DIRETORIA DE GRADUAÇÃO E EDUCAÇÃO PROFISSIONAL
DEPARTAMENTO DE COMPUTAÇÃO
PROGRAMA DE GRADUAÇÃO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

THIAGO RIBEIRO TORACIO

RECONHECIMENTO DE PADRÕES POR MEIO DE FLORESTA DE CAMINHOS ÓTIMOS

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2016

THIAGO RIBEIRO TORACIO

RECONHECIMENTO DE PADRÕES POR MEIO DE FLORESTA DE CAMINHOS ÓTIMOS

Orientadora: Profa. Dra. Priscila Tiemi Maeda Saito

CORNÉLIO PROCÓPIO

2016

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio
Diretoria de Graduação e Educação Profissional
Programa de Graduação em Análise e Desenvolvimento de Sistemas

TERMO DE APROVAÇÃO

RECONHECIMENTO DE PADRÕES POR MEIO DE FLORESTA DE CAMINHOS ÓTIMOS

por

Thiago Ribeiro Toracio

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Tecnólogo em Análise e desenvolvimento de sistemas" e aprovado em sua forma final pelo Programa de Graduação em Análise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná.

Cornélio Procópio, 20/06/2016

Profa. Doutora, Priscila Tiemi Maeda Saito
Universidade Tecnológica Federal do Paraná

Prof. Doutor, Pedro Henrique Bugatti
Universidade Tecnológica Federal do Paraná

Prof. Doutor, Silvio Ricardo Rodrigues Sanches
Universidade Tecnológica Federal do Paraná

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso"

Dedico este trabalho aos meu pai Marcelo e Toracio e minha mãe Tatiana de Mello que estiverem a todo momento presentes em minha vida, e me ajudaram a não desistir, seguindo sempre em frente mesmo com os problemas aparecendo, muitas vezes financeiros. Aos amigos que fiz nesta jornada que me mostraram uma nova realidade e histórias de esforço, me motivando ainda mais. A todos Obrigado.

AGRADECIMENTOS

Agradeço a minha orientadora Priscila Tiemi Maeda Saito que me instruiu e mostrou-me os melhores caminhos na condução desta pesquisa, e a minha querida faculdade Universidade Tecnológica Federal do Paraná, que tem um suporte incrível tanto para seus alunos quanto para seus professores.

RESUMO

TORACIO, THIAGO. RECONHECIMENTO DE PADRÕES POR MEIO DE FLORESTA DE CAMINHOS ÓTIMOS. 26 f. Trabalho de Conclusão de Curso – Programa de Graduação em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2016.

Atualmente existem grandes bases de dados disponíveis, devido aos avanços das tecnologias de aquisição e armazenamento dessas informações. No entanto, há uma grande quantidade de dados não rotulados em relação a uma pequena parte rotulada, tornando-se necessárias técnicas de aprendizado eficazes e eficientes para manipulação e análise dessas informações. Para o aprendizado, é necessário o reconhecimento de determinados padrões, os quais podem ser obtidos por descritores de imagens, que extraem propriedades visuais relacionadas à cor, forma e textura. Algumas características extraídas das imagens podem ser redundantes, outras são mais relevantes na discriminação das imagens. Por isto, após a extração das características das imagens, é importante a análise e a obtenção do vetor de características que melhor descreve o conjunto de dados, aplicando técnicas de redução de dimensionalidade, otimização ou normalizações. Em seguida, podem ser utilizados diferentes procedimentos (supervisionados, não supervisionados e semi-supervisionados) de aprendizado. Este trabalho tem como objetivo o estudo e a análise de técnicas mais efetivas e eficientes para descrição e classificação de bioimagens.

Palavras-chave: reconhecimento de padrões, floresta de caminhos ótimos, classificação supervisionada, processamento de imagens, extração de características

ABSTRACT

TORACIO, THIAGO. PATTERN RECOGNITION THROUGH OPTIMUM-PATH FOREST. 26 f. Trabalho de Conclusão de Curso – Programa de Graduação em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2016.

Currently there are large databases available, due to the advances of the acquisition and storage of this information technologies. However, there is a large amount of unlabeled data in relation to a small section labeled. Becoming necessary effective and efficient learning techniques for manipulation and analysis of this information. For learning the recognition of certain patterns is needed, which can be obtained by imaging descriptors, extracting visual properties related to color, form and texture. Some of the images extracted features may be redundant, others are more relevant to the discrimination of the images. Therefore, after the extraction of the characteristics of images, it is important to analyze and obtain the feature vector that best describes the data set by applying dimensional reduction, optimization and normalization techniques. Then, different procedures may be used (supervised, semi-unsupervised and supervised) learning. This work aims to study and the analysis of more effective and efficient techniques for description and classification of bioimages.

Keywords: pattern recognition, optmum-path forest, supervised classification, image processing, feature extraction

SUMÁRIO

1 INTRODUÇÃO

A quantidade de dados (imagens, vídeos ou textos), disponíveis em diversas áreas da tecnologia, principalmente na Internet, aumenta cada vez mais, formando grandes bases de dados. Surge a necessidade de técnicas para o processamento e a classificação desses dados, de forma a facilitar a manipulação e a análise de informações (HAN et al., 2001). Para diminuir o tempo das buscas e melhorar os resultados são implementados algoritmos que precisam ser melhorados constantemente. Uma área que trata desse estudo é o reconhecimento de padrões (RP).

A descrição do conteúdo semântico dos dados por um ou mais rótulos (palavras-chaves) é a forma mais eficiente e direta de acesso à informação. Este projeto tem como foco a manipulação de *imagens*, sendo atribuídos a cada uma delas um único rótulo. O processo de atribuição de rótulos ocorre de acordo com um determinado número de categorias (classes) referente ao contexto de uma aplicação.

A atribuição manual de rótulos a cada uma das amostras da base de dados torna-se inviável, considerando, principalmente, aplicações de grande porte. Além disso, a fim de evitar erros, a aplicação pode requerer a participação de múltiplos especialistas neste processo de atribuição de rótulos. No entanto, diferentes especialistas podem fornecer descrições distintas, i.e. as atribuições são subjetivas, o que pode gerar inconsistências e resultados insatisfatórios.

Neste contexto, pesquisas têm sido realizadas explorando métodos de classificação automáticos para atribuição dos rótulos. Para tanto, são necessários a identificação/extração de padrões (características) e o treinamento de um modelo (classificador), capaz de rotular automaticamente a base de dados, a partir destes padrões.

Padrões podem ser obtidos por um descritor de imagens, responsável por extrair propriedades visuais (cor, forma e textura) das imagens e armazená-las em vetores de características. Dado dois vetores de características, o descritor compara-os e retorna um valor de distância. Tal valor quantifica a diferença entre as imagens representadas pelos vetores em uma base de dados, a qual pode estar totalmente, parcialmente ou não rotulada. Dependendo da quantidade de

informação disponível dessa base de dados, podem ser aplicados três tipos de técnicas para reconhecimento desses padrões: supervisionadas, semi-supervisionadas ou não-supervisionadas.

1.1 PROBLEMATIZAÇÃO

Um descritor de imagens pode ser classificado dependendo do tipo de informação visual por ele considerado, dentre eles: cor, textura, forma e relacionamento espacial. Trabalhos na literatura Penatti (2009) indicam que não há um descritor com bom desempenho em todas as classes de uma base de dados, bem como em diferentes domínios de aplicação.

Dessa forma, faz-se necessário o estudo comparativo de diferentes descritores para o domínio de bioimagens, foco deste trabalho, bem como a investigação de técnicas que viabilizem a combinação dos melhores descritores de maneira que se possa melhorar a eficácia. Uma das técnicas que poderiam ser utilizadas seria a Programação Genética Torres et al. (2009), compondo vetores de características de alta dimensionalidade para descrever atributos discriminativos em imagens.

No entanto, a alta dimensionalidade de um vetor de características implica custos consideráveis em termos de tempo computacional e requisitos de armazenamento, afetando o desempenho de várias tarefas que utilizam descritores de características, tais como recuperação e classificação de imagens. Para resolver esses problemas, técnicas de redução de dimensionalidade podem também ser investigadas.

A partir das características obtidas, um modelo de classificação pode ser utilizado para rotular automaticamente a base de dados. Apesar de alguns esforços e resultados bem sucedidos em relação à acurácia de classificação, técnicas devem ser melhor investigados em termos de eficiência, principalmente em se tratando de grandes bases de dados. Após a obtenção dos atributos que melhor descrevem as imagens, verifica-se a necessidade da utilização de um modelo de classificação robusto. Apesar de alguns esforços, técnicas mais eficazes e eficientes devem ser melhor investigadas para o domínio de aplicação específico, de forma a apresentar acurácias elevadas e melhor desempenho computacional.

1.2 JUSTIFICATIVA

Devido à crescente quantidade de imagens disponíveis atualmente, torna-se extremamente importante a utilização de mecanismos eficazes e eficientes para a manipulação, análise e recuperação de informações.

Inicialmente, é fundamental a representação das imagens, por meio de suas propriedades visuais, como cor, textura e forma dos objetos. Sendo assim, descritores de imagens, responsáveis pela caracterização das propriedades visuais e pela comparação entre elas, devem ser analisados para o domínio de bioimagens, foco deste trabalho.

O classificador de padrões baseado em floresta de caminhos ótimos (*Optimum-Path Forest* - OPF) Papa (2008), atualmente, tem apresentado bom desempenho tanto em acurácia como em tempo computacional em diferentes domínios de aplicação. O OPF apresenta três versões: supervisionada, não supervisionada e semi-supervisionada, sendo a primeira a mais difundida e utilizada. O OPF foi criado com o intuito de aliar eficiência no processo de treinamento, com eficácia na etapa de classificação dos dados. Essa abordagem apresenta vários benefícios se comparado a outros métodos de classificação de padrões supervisionados: é livre de parâmetros; possui tratamento nativo de problemas multi-classes; e não faz alusão sobre forma e/ou separabilidade das classes, sendo portanto utilizado neste trabalho.

1.3 OBJETIVOS

1.3.1 OBJETIVO GERAL

Este trabalho tem como objetivo realizar um estudo comparativo entre diferentes classificadores, analisando eficácia e eficiência em bioimagens.

1.3.2 OBJETIVOS ESPECÍFICOS

- Verificar diferentes conjuntos de imagens;
- Analisar as características extraídas dos conjuntos de imagens;
- Avaliar estratégias para normalização das características;
- Analisar o desempenho (eficácia e eficiência) de diferentes modelos de classificação.

1.4 ORGANIZAÇÃO DO TEXTO

Este trabalho apresenta a seguinte organização: no capítulo 2 são introduzidos os principais conceitos necessários para a compreensão do projeto proposto, envolvendo reconhecimento de padrões e extração de características, bem como os principais trabalhos existentes

relacionados a este projeto. No capítulo 3 é descrita a metodologia aplicada, incluindo os experimentos realizados e os resultados obtidos. Por fim, no capítulo 4 são apresentadas algumas considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 RECONHECIMENTO DE PADRÕES

Padrão, em reconhecimento de imagens, refere-se a qualquer elemento que possa ser definido quantitativamente mesmo que sujeito a variações. O reconhecimento pode ser realizado por diferenciação ou por classificação (ou ambos). O processo de reconhecimento ou classificação atribui um identificador ou rótulo aos objetos da imagem, baseado nas características providas pelos seus descritores.

2.2 EXTRAÇÃO DE CARACTERÍSTICAS

O processo e representação e descrição visa à extração de características ou propriedades que possam ser utilizadas na discriminação entre classes de objetos. Essas características são, em geral, descritas por atributos numéricos que formam um vetor de características, contendo propriedades visuais, como cor, forma e textura (PENATTI, 2009).

2.2.1 COR

É uma das principais características identificadas pelo olho humano, faz parte de um fenômeno físico-químico em que o comprimento da onda define a cor. É muito difundida e estudada, sendo muito importante na definição da imagem. Para a obtenção das melhores características de cor, atualmente existem quatro principais tipos de descritores: o global, baseado em regiões fixas, o de segmentação e o local (PENATTI, 2009).

2.2.2 FORMA

É um dos meios de descrição, possuindo boa parte da informação do objeto. Características que representam a forma de uma imagem ou vídeo normalmente são extraídas em 2D, neste caso as duas melhores técnicas para extração são as baseadas em região e con-

torno(BOBER, 2001).

2.2.3 TEXTURA

É basicamente um conjunto de determinados padrões definidos pela propriedade dos *pixels* e a relação entre eles, sendo assim, para obtenção de informações relevantes, é necessária a análise global dos elementos desta estrutura. No conceito de imagens digitais, normalmente são utilizadas três abordagens características: estatística, estrutural e a espectral Gonzales e Woods (2002) para definir uma textura.

2.3 CLASSIFICAÇÃO DE DADOS

Apresenta como foco o desenvolvimento de técnicas computacionais sobre o aprendizado e a construção de sistemas que permitam adquirir conhecimento automaticamente. Existem três técnicas (PAPA, 2008), sendo elas:

- Supervisionado: são identificados a priori o padrão e as classes, adquirindo amostras representativas para cada uma das classes que se deseja classificar. Exemplos de classificadores supervisionados: Support Vector Machines (SVM) Tong et al. (2002) e o Optimum-Path Forest (OPF).
- Semi-supervisionado: é um híbrido entre o supervisionado e o não-supervisionado, diminuindo a necessidade de dados rotulados, quando somente um pequeno conjunto de exemplos rotulados está disponível. Exemplos de classificadores semi-supervisionados: SemiL e o OPF.
- Não-supervisionado: não se conhece o padrão, nem o número total de classes a serem encontradas durante a classificação. Também conhecido como análise de agrupamentos (clusters), o conjunto de dados é particionado em grupos, baseados em características específicas, os dados são separados em grupos distintos onde as amostras similares se agrupam. Exemplos de classificadores não supervisionados: K-means e OPF.

2.4 CLASSIFICAÇÃO BASEADA EM FLORESTA DE CAMINHOS ÓTIMOS

2.4.1 TREINAMENTO

Nesta fase há o treinamento de amostras de classes distintas com o objetivo de encontrar um padrão para que o classificador consiga de maneira eficaz rotular novas amostras.

Esses objetos são representados por um vetor de características. As amostras de treinamento são interpretadas como os nós de um grafo direcionado, cujos arcos são definidos através de uma relação de adjacência e ponderados por uma função de distância. Espera-se que objetos da mesma classe estejam conectados através de suas amostras. Em seguida, encontra a amostra mais representativa, que é chamada de protótipo ou raiz da árvore.

2.4.2 AVALIAÇÃO

Em muitos conjuntos de dados há redundância e os objetos escolhidos para treinamento acabam por não representar as classes da melhor forma. Um meio de atenuar esse problema é selecionar amostras mais representativas. O OPF oferece um passo de avaliação, que utiliza um conjunto de dados rotulados chamado conjunto de avaliação, e iterativamente calcula acurácia do conjunto de treinamento. A cada iteração permite trocar dados entre o conjunto de treinamento e avaliação, com o objetivo de selecionar objetos mais representativos. Não há mudança no tamanho do arquivo de treinamento.

2.4.3 TESTE

Após o treinamento do classificador, obtém-se uma floresta de caminhos ótimos, na qual cada árvore representa uma classe e cada classe possui um protótipo, o elemento mais representativo de cada classe. Dessa maneira, quando uma nova amostra é adicionada no espaço de características, considera-se esse novo objeto como parte original do grafo. A classificação consiste em encontrar qual nó de treinamento oferece o caminho ótimo para o novo objeto, em relação ao protótipo de cada árvore. Assim, são considerados todos os possíveis caminhos a partir dos protótipos no grafo de treinamento, extendidos para o novo objeto por um arco. O novo objeto é rotulado com a classe que oferece esse caminho ótimo.

2.5 TRABALHOS RELACIONADOS

Segundo Lorena e Carvalho (2006), muitas funções celulares são realizadas em compartimentos específicos da célula. A previsão da localização celular de uma proteína é assim relacionada com a sua identificação. Este trabalho utiliza duas técnicas de Aprendizado de Máquina, *Support Vector Machines* (SVMs), para identificar a localização de proteínas a partir de três categorias de organismos: bactérias gram-positivas e gram-negativas e fungos. A tarefa de localização tem várias classes, as quais correspondem aos possíveis locais de proteína. Além disso, este trabalho também investiga e compara várias estratégias para estender esta técnica

para executar previsões de múltiplas classes.

Em Mansano et al. (2012) é retratado que a eficiência nas tarefas de classificação de imagem podem ser melhoradas por meio de informações fornecidas por várias fontes combinadas, tais como forma, cor, textura e as propriedades visuais. Esta robustez da técnica é avaliada pela Floresta de Caminhos Ótimos. Os experimentos mostraram que a metodologia proposta pode superar facilmente a informação individual fornecida por descritores únicos.

Segundo Santos et al. (2012), um enorme esforço tem sido aplicado na classificação de imagens para criar mapas temáticos de alta qualidade e de estabelecer inventários mais precisos sobre o melhor uso da superfície do solo com imagens de sensoriamento remoto (LER). Este trabalho propôs um classificador adaptado para escalas múltiplas de segmentação, através da combinação de classificadores fracos para construir um mais eficiente. Os experimentos foram realizados em imagens de plantações de café. É mostrado ainda que a abordagem pode detectar em uma escala muito maior o conjunto de características que melhor se adaptem às necessidades do cliente.

Segundo Faria et al. (2014), o crescimento frequente de dados visual, seja por imagens ou vídeos têm contribuído enormemente para a chamada *'big-data revolution'*. Este montante de corte de dados visuais dá origem a diversos novos problemas de classificação visual nunca antes imaginados. O trabalho combina métodos de caracterização de imagem e de aprendizagem através de uma abordagem de meta-aprendizagem responsável por avaliar quais métodos são os melhores para auxiliar na solução de um determinado problema. A estrutura usa uma estratégia de seleção de classificadores menos correlacionadas, mas eficazes, através de uma série de medidas de análises da diversidade. os experimentos mostram que a abordagem proposta alcança resultados comparáveis aos algoritmos bem conhecidos da literatura sobre quatro aplicações diferentes, mas utilizando menos métodos de aprendizagem e descrição.

Segundo Souza et al. (2005), estão surgindo tecnologias que permitem aos biólogos compreender melhor as interações entre vários estados patológicos a nível de genes como *microarray*(vetores pequenos). No entanto, a quantidade de dados gerados por estas ferramentas torna-se problemática quando os dados devem ser analisados sem interação humana, ou seja, automaticamente. Neste trabalho, os autores apresentaram um novo método de seleção para genes baseado em Algoritmos Genéticos e *Support Vector Machines* (SVM) com intuito de classificar as amostras de tecido. Para tal, os autores utilizaram uma estimativa de erro no SVM para avaliar cada indivíduo. O método proposto foi comparado com técnicas comuns de seleção. Os resultados experimentais utilizando conjuntos de dados públicos de *microarray* demonstraram bons resultados.

Segundo Muhammed et al. (2012), nos últimos anos vem crescendo a busca por anotações de imagens automáticas (AIA), essa técnica é utilizada para facilitar a extração de alto nível das características, a partir de imagens por meio de técnicas de aprendizado de máquina. Muitas técnicas AIA utilizam a análise de recurso como o primeiro passo para identificar os objetos na imagem. No entanto, a alta dimensionalidade das características da imagem tornam o desempenho do sistema fraco. O trabalho descreve e avalia uma anotação automática de imagem *framework* que utiliza descritores SURF para selecionar o número certo de características e recursos adequados para anotação, com uma estrutura híbrida, em que *k-means clustering classification* é usado na fase de treinamento e o classificador *fuzzy K-NN* na fase de anotação.

O trabalho de Faria et al. (2010) apresenta a ideia de combinação de descritores por Enxame de Partículas e suas aplicações. Para efeitos de classificação foi utilizada a técnica *Optimum-Path Forest* (OPF), que interpreta as amostras do conjunto de dados, como os nós de um grafo. O método combina diferentes descritores de cor para a classificação, utilizando o conjunto de dados COREL e, ainda, os arcos do grafo do OPF por serem semelhantes. Esta abordagem proporciona melhores níveis de precisão, do que, cada descritor individualmente.

Barros et al. (2012) apresenta um levantamento de algoritmos evolucionários que são projetados para a utilização de *decision-tree* (arvore de decisão). Neste contexto, a maior parte do trabalho concentra-se em abordagens que evoluem soluções tradicionais como *top-down* e *divideand-conquer*. Além disso, apresenta algumas alternativas que fazem uso de algoritmos evolutivos para melhorar componentes de classificadores *decision-tree* e facilita a utilização desses algoritmos em diferentes domínios.

Em Lorena et al. (2011), diferentemente dos outros trabalhos relacionados, é proposta uma abordagem para a seleção de atributos em conjuntos de dados com apenas uma classe. A proposta baseia-se na combinação de diferentes medidas de atributos adaptadas para o cenário de uma única classe, com a intenção de que ela possa identificar quais atributos fazem parte da classe e quais não, o que facilita o processo já que não irá precisar de diversas combinações de classes.

O artigo de Kimura1 et al. (2011) apresenta uma avaliação de descritores de imagens para pesquisas em grandes bancos de dados, utilizando vários descritores de imagens propostos na literatura, os resultados obtidos mostram que em geral, a eficácia de recuperação dos diferentes descritores varia pouco em coleções de imagens pequenas, enquanto que em uma grande coleção há uma boa diferença entre a eficiência e o tempo de processo. Além disso, propõe dois descritores para serem usados em bancos de dados grandes e heterogêneos de imagens.

3 METODOLOGIA DE PESQUISA

O foco deste trabalho é a análise e comparação do desempenho de diferentes classificadores em bioimagens. Os conjuntos de dados e as ferramentas e classificadores utilizados são descritos nas seções 3.1 e 3.2

3.1 CONJUNTOS DE DADOS

Para os experimentos foram considerados 7 conjuntos médicos disponíveis em 'http://datam.i2r.a-star.edu.sg/datasets/krbd/' que é um repositório online com conjuntos de alta dimensionalidade biomédica. Seguem as respectivas descrições para cada um dos conjuntos.

- **ALL-AML Leukemia:** contém 72 amostras, descritas por 7.129 atributos numéricos e distribuídos em duas classes distintas. As medições são correspondentes às amostras ALL (leucemia linfocítica aguda) e AML (leucemia mieloide aguda) da medula óssea e sangue periférico.

- **Breast Cancer:** composto por 97 amostras e 24.481 atributos, representativas de resultados de pacientes com câncer de mama. Os dados correspondem a pacientes que desenvolveram metástases à distância no prazo de 5 anos, rotulados como (recaída) e, pacientes que permaneceram saudáveis da doença após o diagnóstico inicial por um intervalo de pelo menos 5 anos, rotulado como (não-recaída).

- **Central Nervous System Embryonal Tumor**: contém 60 amostras e 7.129 atributos, referentes a 21 pacientes sobreviventes ao tratamento (classe 1) e 39 pacientes que falaceram (classe 2).

- **Colon Tumor:** composto por 62 amostras coletadas de pacientes com câncer no cólon. Entre eles, 40 biópsias de tumor são de tumores referidas como negativa e 22 rotuladas como positiva.

Tabela 1: Valores de cada Conjunto

	Amostras	Classes	Atributos
ALL-AML Leukemia	72	2	7.129
Breast Cancer	97	2	24.481
Central Nervous System	60	2	7.129
Colon Tumor Dataset	62	2	2.000
MLL Leukemia	72	3	12.582
Ovarian Cancer	253	2	15.154
Prostate Tumor	136	2	12.600

- **MLL Leukemia:** consiste de 72 amostras, sendo 24 linfoblástica aguda (LLA), 20 leucemia de linhagem mista (MLL) e 28 leucemia mielóide aguda (LMA).
- **Ovarian Cancer:** contém 253 amostras referentes a espectros *proteomics* gerados por espectroscopia de massa. O conjunto é composto por 91 controles rotuladas como normal e 162 como câncer de ovário. Os dados encontram-se normalizados por meio da estratégia de normalização MinMax.
- **Prostate Cancer:** consiste em 136 amostras de pacientes, sendo 77 rotuladas como tumor e 59 como normal.

3.2 FERRAMENTAS E FRAMEWORKS

Neste capítulo são descritos as ferramentas e frameworks utilizados nos experimentos, de forma a verificar o desempenho de cada classificador aplicado aos conjuntos de imagens descritos anteriormente, bem como a descrição dos classificadores considerados nos experimentos.

- **LIBOPF:** framework baseado em Floresta de Caminhos Ótimos, conceituado por seu baixo tempo de processamento em relação a todo o processo de classificação, bem como por sua acurácia elevada.
- **WEKA:** *software* de código aberto emitido sob a GNU General Public License (GPL), sendo um conjunto de algoritmos para o aprendizado de máquina, que contém ferramentas para pré-processamento de dados, classificação, regressão, clustering, regras de associação, e visualização. Sendo escrito na linguagem Java e contém uma GUI que interage com os arquivos de dados para gerar resultados visuais. Outro fator interessante

é que ele possui uma API geral, sendo possível incorporá-lo, como qualquer outra biblioteca, aos seus aplicativos e, possui uma interface gráfica que pode ser utilizada para realizar os treinamentos e testes dos classificadores. Para a realização dos experimentos foram utilizados alguns classifidores que estão presentes nesta ferramenta.

- **NaiveBayes (NB):** classificador de aprendizagem supervisionada probabilístico baseado no teorema de bayes, sendo um dos melhores e mais simples da atualidade (ZHANG, 2004).

- **IBK:** algoritmo de aprendizagem baseado em instância. Esse tipo de algoritmo é derivado do método de classificação *k-nearest neighbor* (KNN). Com a diferença de que o algoritmo do tipo IBK é incremental e tem como foco maximizar a acurácia sobre novas instâncias do problema (AHA et al., 1991).

- **J48:** derivado do algoritmo C4.5, atualmente implementado em java. Sua finalidade é gerar uma árvore de decisão baseada em um conjunto de dados de treinamento. Para a montagem da árvore, o algoritmo J48 utiliza a abordagem de dividir um problema complexo em subproblemas mais simples e, assim, aplicando recursivamente a mesma estratégia para cada parte do problema, dividindo o espaço definido pelos atributos em subespaços, associando para cada um deles uma classe. (WITTEN; FRANK, 2005).

3.3 EXPERIMENTOS

Os experimentos foram realizados em um notebook da Dell inspiron 14r no sistema operacional aberto ubuntu, sendo este uma derivação do linux.

Para comparação entre os classificadores, foram obtidos os conjuntos de características que estavam em formato txt. Tais conjuntos foram transformados em formato dat, que é o tipo de arquivo aceito como entrada para o OPF. Com o conjunto devidamente transformado, aplicou-se a divisão (“split”) do arquivo separando o mesmo em 80% de treinamento e em 20% de teste. Foram realizadas 10 execuções para cada um dos classificadores e conjuntos de dados. Para executar os experimentos no Weka foi realizada a conversão dos arquivos de treinamento e teste em formato arff, mantendo as mesmas amostras utilizadas nos experimentos com o OPF, para fins de comparação de desempenho entre os classificadores. Após todos os arquivos devidamente transformados, o próximo passo são as execuções com todos os classificadores. As Figuras 1 e 2 mostram como foram realizadas as execuções no Weka. Primeiramente, é necessário clicar no botão “explorer”, o qual irá abrir outra tela, que será o “Preprocess”, em que é necessário abrir o arquivo de treinamento para dar continuidade ao experimento.

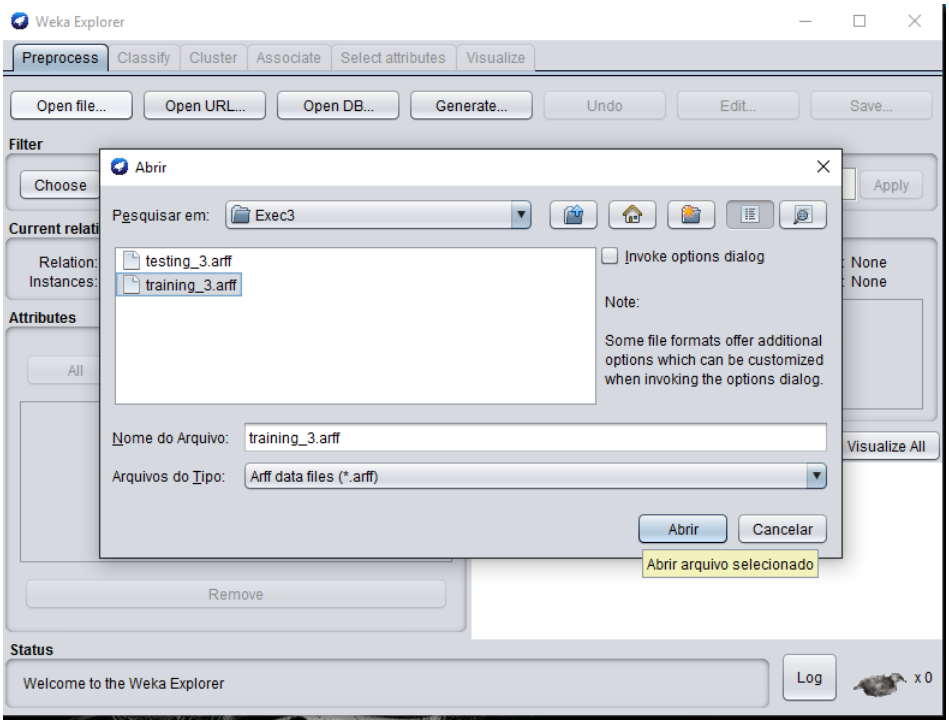

Figura 1: WEKA - Selecionando o arquivo de treinamento

Após o carregamento do arquivo de treinamento, marca-se a opção “use train set” e realiza-se o treinamento. Em seguida, carrega-se o arquivo de teste, aplica-se a opção “Supplied teste set” e executa-se o teste.

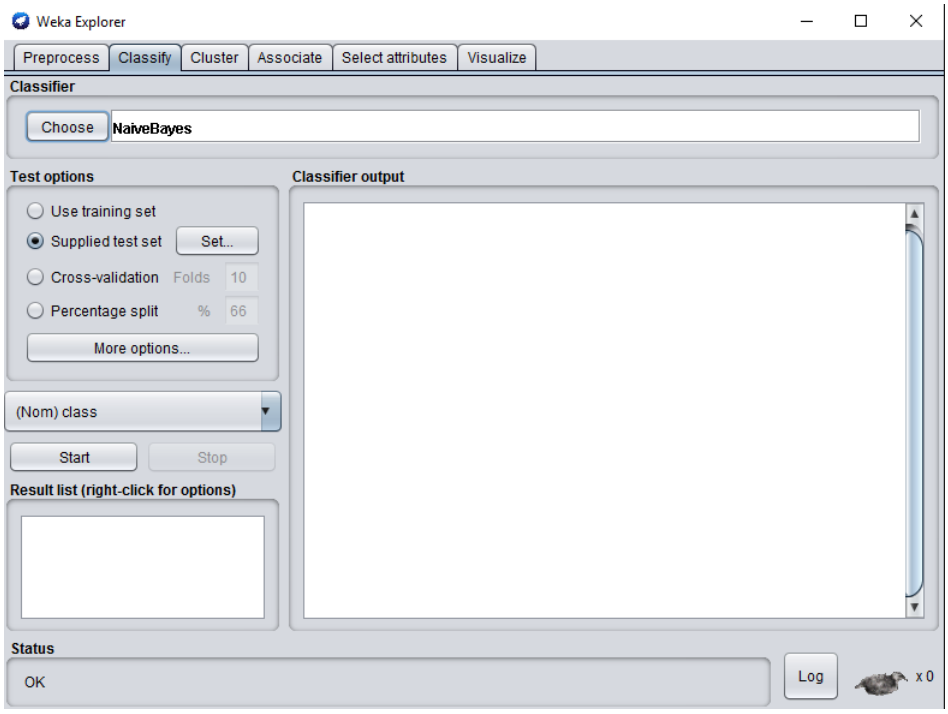

Figura 2: WEKA - Escolhendo Supplied teste set

Após o fim da execução, gera-se um arquivo com os dados, contendo os valores de

acurácia e tempo de teste, os quais foram armazenados para as comparações entre classificadores. As execuções no OPF seguiram o mesmo padrão das execuções no WEKA, porém em linhas de comando por meio do terminal do Ubuntu, utilizando o treinamento, a classificação e obtendo a acurácia dos arquivos de treinamento e de teste. No entanto, no OPF são gerados vários arquivos separados com tempo e precisão, no Weka é gerado apenas 1 arquivo com todos os valores.

4 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados por meio dos experimentos realizados referentes às acurácias e tempo computacional de treinamento e de teste (Tabelas 2-4, respectivamente) obtidos por cada classificador (OPF, IBK, J48 e NB) para cada um dos conjuntos analisados (ALL-AML Leukemia, Breast Cancer, Central Nervous System, Colon Tumor Dataset, MLL Leukemia, Ovarian Cancer, Prostate Tumor).

Tabela 2: Média de Acurácia

	OPF	IBK	J48	NB
ALL-AML Leukemia	**87,00** $\pm 6,32$	**88,00** $\pm 4,22$	**84,66** $\pm 5,49$	$80,28 \pm 6,20$
Breast Cancer	**61,77** $\pm 7,90$	**58,57** $\pm 9,27$	**58,56** $\pm 12,10$	$54,76 \pm 3,37$
Central Nervous System	**58,25** $\pm 12,18$	**55,38** $\pm 10,13$	**60,77** $\pm 10,54$	**63,07** $\pm 13,95$
Colon Tumor Dataset	**75,63** $\pm 7,00$	**69,23** $\pm 11,47$	**72,30** $\pm 16,30$	$55,48 \pm 10,03$
MLL Leukemia	**96,86** $\pm 4,35$	$77,33 \pm 7,83$	$84,00 \pm 7,17$	**92,66** $\pm 6,63$
Ovarian Cancer	$92,70 \pm 2,12$	$94,03 \pm 1,69$	**96,34** $\pm 3,20$	**94,03** $\pm 3,33$
Prostate Tumor	**78,45** $\pm 5,70$	**75,00** $\pm 8,59$	**78,57** $\pm 4,76$	$54,64 \pm 8,26$

Tabela 3: Tempo computacional de Treinamento em segundos

	OPF	IBK	J48	NB
ALL-AML Leukemia	$0,032 \pm 0,0098$	$0,183 \pm 0,0495$	**0,011** $\pm 0,0032$	$0,373 \pm 0,1601$
Breast Cancer	$0,145 \pm 0,0118$	$1,060 \pm 0,1865$	**0,026** $\pm 0,0506$	$1,377 \pm 0,3458$
Central Nervous System	$0,020 \pm 0,0081$	$0,126 \pm 0,0051$	**0,001** $\pm 0,0000$	$0,188 \pm 0,0382$
Colon Tumor Dataset	$0,008 \pm 0,0004$	$0,036 \pm 0,0052$	**0,001** $\pm 0,0000$	$0,059 \pm 0,0088$
MLL Leukemia	$0,054 \pm 0,0116$	$0,307 \pm 0,0507$	**0,013** $\pm 0,0095$	$0,604 \pm 0,6555$
Ovarian Cancer	$0,695 \pm 0,0065$	$1,834 \pm 0,0695$	**0,034** $\pm 0,0236$	$1,759 \pm 0,2907$
Prostate Tumor	$0,170 \pm 0,0101$	$0,555 \pm 0,0212$	**0,024** $\pm 0,0443$	$0,738 \pm 0,2230$

Tabela 4: Tempo computacional de Teste em segundos

	OPF	IBK	J48	NB
ALL-AML Leukemia	**0,005** $\pm 0,0010$	$0,122 \pm 0,0181$	$0,033 \pm 0,0116$	$0,471 \pm 0,1467$
Breast Cancer	**0,035** $\pm 0,0025$	$0,590 \pm 0,0485$	$0,119 \pm 0,0247$	$0,695 \pm 0,2439$
Central Nervous System	**0,004** $\pm 0,0003$	$0,151 \pm 0,0726$	$0,029 \pm 0,0137$	$0,132 \pm 0,1085$
Colon Tumor Dataset	**0,001** $\pm 0,0003$	$0,039 \pm 0,0087$	$0,015 \pm 0,0084$	$0,024 \pm 0,0127$
MLL Leukemia	**0,007** $\pm 0,0018$	$0,270 \pm 0,0394$	$0,086 \pm 0,0254$	$0,299 \pm 0,1211$
Ovarian Cancer	**0,111** $\pm 0,0063$	$0,998 \pm 0,0868$	$0,183 \pm 0,0245$	$0,820 \pm 0,2677$
Prostate Tumor	**0,031** $\pm 0,0037$	$0,326 \pm 0,0246$	$0,183 \pm 0,0245$	$0,294 \pm 0,0360$

Analisando as acurácias obtidas, os classificarores OPF, IBK e J48 apresentarem os melhores resultados, sendo equivalentes, conforme as médias e desvios da Tabela 1. Com relação ao tempo computacional para treinamento, os classificadores J48 e OPF apresentaram um melhor resultado. Já para o tempo computacional para teste, o classificador OPF obteve melhores resultados para todos os conjuntos de dados analisados.

O conjunto Ovarian Cancer obteve melhores resultados (maior média de acurácia e menor desvio padrão) em relação aos demais conjuntos. Vale destacar que o conjunto Ovarian Cancer encontra-se com as características normalizadas por meio da estratégia de normalização Min-Max.

5 CONCLUSÃO

Por meio dos resultados obtidos, nota-se que, em relação à acurácia, os classificadores OPF, IBK e J48 apresentaram um desempenho equivalente em todos os conjuntos analisados. Com relação ao tempo computacional, os classificadores J48 e OPF apresentarem menores tempos para treinamento, e o OPF apresentou menor tempo para teste para todos os conjuntos analisados. Vale ressaltar, que cada classificador pode apresentar um melhor desempenho, dependendo do tipo de conjunto em que seja aplicado. Além disso, melhorias nos resultados podem ser obtidas por meio de aplicação de algumas estratégias de normalização para as características dos conjuntos de dados, bem como estratégias de otimização ou de seleção de características.

REFERÊNCIAS

AHA, D. W.; KIBLER, D.; ALBER, M. K. **Instance-Based Learning Algorithms**. 1991. 40–66 p.

BARROS, R. C. et al. **A Survey of Evolutionary Algorithms for Decision-Tree Induction**. 2012. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5928432>.

BOBER, M. **Visual Shape Descriptors**. 2001.

FARIA, F. A. et al. **Multimodal Pattern Recognition Through Particle Swarm Optimization**. 2010.

FARIA, F. A. et al. **A framework for selection and fusion of pattern classifiers in multimedia recognition**. 2014.

GONZALES, R. C.; WOODS, R. E. **Digital Image Processing**. 2002.

HAN, J.; KAMBER, M.; PEI, J. **Concepts and Techniques**. 2001. Disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=000438287>. Acesso em: 13 de novembro de 2014.

KIMURA1, P. A. S. et al. **Evaluating Retrieval Effectiveness of Descriptors for Searching in Large Image Databases**. 2011.

LORENA, A. C.; CARVALHO, A. C. de. **Protein cellular localization prediction with Support Vector Machines and Decision Trees**. 2006.

LORENA, L. H. N.; CARVALHO, A. C. P. L. F.; LORENA, A. C. **Selecao de atributos em problemas de classificacao com uma unica classe**. 2011. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0055.pdf>.

MANSANO, A. et al. **Improving Image Classification Through Descriptor Combination**. 2012.

MUHAMMED, V.; KUMAR, G. S.; SREERAJ. **Automatic Image Annotation Using SURF Descriptors**. 2012.

PAPA, J. P. **Classificação supervisionada de padrões utilizando floresta de caminhos ótimos**. 2008. Universidade Estadual de Campinas - Instituto de Computação. Disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=000438287>.

PENATTI, O. A. B. **Estudo Comparativo de Descritores para Recuperacao de Imagens por Conteudo na Web**. 2009. Universidade Estadual de Campinas - Instituto de Computação. Disponível em: <http://www.recod.ic.unicamp.br/ otavio/academico/files/dissertacaoFinal.pdf>.

SANTOS, J. A. dos et al. **Multiscale Classification of Remote Sensing Images**. 2012.

SOUZA, B. F. de; CARVALHO, A. C. P.; TICONA, W. C. **Applying genetic algorithms and SVMs to the gene selection**. 2005.

TONG, S.; KOLLER; DAPHNE. **Support vector machine active learning with applications to text classification**. [S.l.]: JMLR, 2002. 45–66 p.

TORRES, R. da S. et al. **A genetic programming framework for content-based image retrieval**. 2009. 283-292 p.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2005.

ZHANG, H. The optimality of naive bayes. 2004. Disponível em: <http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf>.