Estudo e Implementação de Algoritmos de Agrupamento e de Rotulação Aplicados no Diagnóstico por Imagens de Patologias Renais

Arthur C. Basílio[1], Pedro Antonio F. da Silva[1], Vinícius P. Machado[1], Nayze Lucema S. Aldeman[2]

[1]Departamento de Ciência da Computação – Universidade Federal do Piauí (UFPI) – Teresina, PI – Brasil

[2]Departamento de Medicina Especializada – Universidade Federal do Piauí (UFPI) – Teresina, PI - Brasil

basilio.arth@gmail.com,p.antonio.f.s@gmail.com,vinicius@ufpi.edu.br, nayzealdeman@gmail.com

Abstract: This article proposes the application of clustering and labeling algorithms to aid in the diagnosis of renal pathologies. The information is extracted from the renal images provided in conjunction with the database containing nephrology diagnostic records.

Through unsupervised learning algorithms, different groups are formed. Through the labeling algorithm, the main attributes that characterize each group are revealed, facilitating their understanding. The article also demonstrates the influence of such information extracted from the images, as well as the aid in the knowledge acquisition process promoted by the grouping and labeling.

Resumo: Este artigo propõe a aplicação de algoritmos de agrupamento e rotulação para o auxílio no diagnóstico de patologias renais. As informações são extraídas das imagens renais fornecidas em conjunto com a base de dados contendo registros de diagnóstico de nefrologias.

Através dos algoritmos de aprendizagem não supervisionada, diferentes grupos são formados. Por intermédio do algoritmo rotulador, os principais atributos que caracterizam cada grupo se revelam, facilitando sua compreensão. O artigo demonstra também a influência de tais informações extraídas das imagens, bem como o auxílio no processo de aquisição de conhecimento promovido pelo agrupamento e rotulação.

1. Introdução

Por meios convencionais, o diagnóstico das patologias renais é dado através da análise particular de cada imagem por um nefrologista (médico especializado no diagnóstico e tratamento clínico de doenças renais). A partir dessa análise, o nefrologista obtém informações relevantes sobre o rim do paciente de acordo com as diferentes características apresentadas na imagem. Portanto, extrair informações relevantes do conjunto de imagens disponibilizado é parte primordial do projeto e tem grande influência no seu resultado.

Dado este contexto, este artigo tem como objetivo apresentar o uso de técnicas de aprendizagem de máquina não supervisionadas para auxiliar o diagnóstico de doenças

renais. Através dos algoritmos de aprendizagem não supervisionada, serão formados agrupamentos baseados nos dados. Com a utilização do algoritmo rotulador, os principais atributos que caracterizam cada grupo se revelam, facilitando sua compreensão e consequentemente diagnósticos de patologias.

2. Revisão de Literatura

Os recursos tecnológicos tendem a abandonar o posto de ferramenta auxiliar para se tornar elemento essencial no avanço da ciência e na consequente promoção de novas descobertas para o bem-estar da sociedade. Para a estruturação das etapas deste trabalho utilizou-se o processo de KDD (do inglês, Knowledge-Discovery in Databases) proposto por Fayyad et al. (1996).

O processo de KDD possui cinco etapas, são elas: seleção, pré-processamento, transformação, mineração de dados e avaliação ou interpretação. Durante a seleção identifica-se quais informações dentre os dados existentes na base de dados devem ser efetivamente consideradas durante o processo de KDD. No pré-processamento, informações inconsistentes, errôneas ou até mesmo inexistentes devem ser corrigidas com o intuito de não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD. A próxima etapa, transformação, trata-se de rotinas aplicadas aos dados para adequá-los aos algoritmos aplicados futuramente, que exigem formatos específicos de entrada ou que melhor desempenham com dados em uma determinada forma. Tornar valores contínuos em faixas discretas e normalização são exemplos de métodos aplicados.

Na mineração de dados, por sua vez, é onde são utilizados os algoritmos de aprendizado de máquina para extrair padrões de dados. Ao final, é o momento de definir métodos para avaliar os resultados encontrados na mineração, pode-se, por exemplo, medir a acurácia de uma classificação. Também pode ser necessário interpretar e tratar os resultados para melhor adequá-los a proposta do estudo. Este trabalho também envolve o processo de extração de informações das imagens disponibilizadas. Utilizou-se como base para o processamento digital de imagens (PDI) alguns dos passos especificados por Gonzalez e Woods (2009).

A aquisição de imagens é o primeiro processo e ele pode ser tão simples quanto receber uma imagem que já esteja em formato digital. Em geral, o estágio de aquisição de informações de imagens envolve um pré-processamento, por exemplo, o redimensionamento de imagens. O realce de imagens é o processo de manipular uma imagem de forma que o resultado seja mais adequado do que o original para uma aplicação específica. Tal etapa estabelece deste o início que as técnicas utilizadas são orientadas de acordo com o problema. Dessa forma, por exemplo, um método bastante útil para realçar imagens radiográficas pode não ser a melhor abordagem para realçar imagens de satélite capturadas na banda infravermelha do espectro eletromagnético.

O processamento morfológico lida com ferramentas para a extração de componentes de imagens úteis na representação e descrição da forma. Os resultados retornados por essa etapa são considerados atributos das imagens. Como dito por Gonzalez e Woods (2009), as técnicas utilizadas tanto no pré-processamento das imagens quanto na aquisição de informações relevantes variam de acordo com o objetivo em questão. Para este trabalho em questão essas três etapas – aquisição, pré-processamento

e processamento morfológico – do processamento digital de imagens foram utilizadas e se encontram especificadas na seção 3.3.2.

Outra tecnologia que iremos utilizar é o aprendizado de máquina (AM), uma subárea da Inteligência Artificial que pode ser descrito como o desenvolvimento de técnicas computacionais sobre o aprendizado e construção de sistemas capazes de adquirir conhecimento de forma automática [Santos 2005]. Este artigo utiliza, dentre os ramos do AM, o aprendizado não supervisionado, que consiste na descoberta das relações implícitas em um conjunto de dados não rotulados [Barber 2012]. Em outras palavras, identifica padrões para agrupar dados.

O rotulador também é parte importante deste trabalho. O processo de rotulação consiste em nomear os grupos formados de acordo com as suas principais características. Essas principais características são concebidas através dos rótulos - atributos que têm forte influência sobre o perfil de um grupo, e o valor mais recorrente desse atributo (informação explicitada na tabela 2 deste artigo). De maneira abrangente, rotular significa apresentar uma identificação clara de cada grupo. [Lucas et al. 2014].

3. Metodologia

O desenvolvimento deste trabalho baseia-se na metodologia descrita por Fayyad et al. (1996), detalhada na seção 2. A seguir detalharemos essas etapas.

3.1. Seleção

A base original disponibilizada para este projeto contém 100 registros com 31 atributos categóricos cada. Desses 31 atributos, 30 descrevem características sobre o rim observado pelo nefrologista responsável e 1 se trata do diagnóstico propriamente dito (também descrito pelo nefrologista). Além da base de dados citada, um conjunto de 71 imagens (lâminas) renais foram disponibilizadas pela mesma fonte. Cada uma dessas imagens possui associação direta e respectiva a um registro da base de dados. Portanto, 29 registros da base não possuem uma imagem renal como referência. Por esse motivo, foi convencionado que tais registros recebam a valor ‘inexistente’ para os atributos correspondentes ao retirados das imagens.

Apesar das circunstâncias, todos os registros foram considerados durante o desenvolver do projeto e nenhum dos atributos já existentes foram eliminados por motivos de irrelevância para a aplicação.

3.2. Transformação

3.2.1 Transformação da Base de Dados

O projeto em questão tem como proposta o uso de algoritmos de aprendizagem de máquina não supervisionados e algoritmos de rotulação. Para o primeiro caso, os seguintes algoritmos foram utilizados: *Agglomerative Clustering*, *Birch Clustering*, *KMeans Clustering* e *MiniBatchKMeans Clustering* [Pedregosa et al. 2011].

Para que os dados contidos na base original fossem compatíveis com os algoritmos escolhidos, foi necessária a transformações dos atributos categóricos para a forma numérica. Tal transformação assume a seguinte forma: cada valor diferente assumido por cada atributo é transformado em uma nova coluna. A coluna que deu origem a essas novas colunas tem seus valores redistribuídos, assumindo o valor 1 quando sua descrição

coincide com o nome da coluna ou 0, caso contrário. Ao fim da transformação a coluna de origem é deletada da base de dados.

3.2.2 Extraindo Atributos das Imagens

Como mencionado anteriormente, as imagens são parte fundamental do projeto, pois são a partir delas que os diagnósticos são realizados pelos especialistas. Por esse motivo, encontrar a melhor maneira de retirar delas informações (extrair descritores das imagens) que sejam compatíveis com a aplicação é um processo de suma importância. Serviram de base para essa etapa os passos descritos por Gonzalez, R. C. (2009) especificados na seção 2, detalhadas a seguir.

As imagens foram adquiridas através da disponibilização de informações por parte dos profissionais da saúde envolvidos com o projeto. Contabiliza-se um total de 71 imagens de lâminas renais. Tais imagens já se encontram em formato digital. Nessa etapa alguns tratamentos foram aplicados às imagens com o objetivo de padronizar a forma como elas serão disponibilizadas aos algoritmos e, consequentemente, torná-las mais adequadas ao uso neste trabalho. Dependendo do aparelho clínico utilizado para a produção das imagens, a cor predominante da imagem varia. Além disso, por também característica do aparelho, algumas legendas eram aplicadas à imagem.

Como forma de contornar as subjetividades de cada aparelho, todas as imagens tiveram sua coloração alterada para escalas de cinza. Além disso, para a retirada das legendas foram aplicadas técnicas de processamento morfológico como abertura e fechamento [Gonzalez e Woods 2009]. A fim de que o método utilizado seja robusto à presença de ruído, a variações na aquisição das imagens e que necessite da mínima intervenção do usuário ou configuração de parâmetros, extrair dados estatísticos das imagens foi a forma escolhida para se obter informações relevantes sobre elas. Tais informações estatísticas foram consideradas como novos atributos da base de dados e, assim como explicado em tópicos futuros, existe um embasamento matemático por trás de cada uma delas. Por esse motivo, caso futuramente uma nova imagem seja adquirida, sobre ela será aplicado o mesmo processo (fórmulas) já submetidas às imagens previamente existentes, obtendo-se novas informações com o mesmo nível de fidelidade.

Além disso, apesar de não existir uma definição formal de textura de uma imagem, as informações estatísticas estão diretamente ligadas com tal atributo. Esse fato reforça o uso desses dados, pois se admitiu que informações sobre a textura dos rins estariam relacionadas com a atual situação em que eles se encontram, consequentemente sendo uma informação valiosa para o diagnóstico.

3.2.3 Atributos

A seguir se encontram todos os dados estatísticos retirados das imagens: *Média*, *Desvio Padrão*, *Descritor de Suavidade Relativa*, *Terceiro Momento*, *Uniformidade* e *Entropia*. Tais atributos, quando analisados isoladamente, podem não ser suficientes para concluir uma informação concreta sobre a textura da imagem. Porém, quando analisadas em conjunto, a textura é devidamente descrita [Weiner et al. 2014].

A média nos diz a intensidade média de cada região da imagem e só é útil como uma ideia aproximada da intensidade, não da intensidade propriamente dita. O desvio padrão relata a variabilidade nos níveis de intensidade da imagem. O descritor de suavidade relativa fornece uma medida de contraste de intensidade que pode ser usada

para estabelecer descrições de variância relativa. O terceiro momento, por sua vez, é uma medida de assimetria do Histograma (relação entre um nível de intensidade existente na imagem e a sua respectiva probabilidade de ocorrência nessa mesma imagem). Isso dá uma ideia aproximada de se os níveis de intensidade tendem para o lado escuro ou claro em torno da média. A uniformidade é o grau de invariabilidade relativa dos níveis de intensidade em toda a extensão da imagem. Por fim, a entropia (ou incerteza) da imagem pode ser definida como um número quantificador da aleatoriedade da imagem, ou seja, quanto maior for esse número, mais irregular, atípica ou não padronizada será a imagem analisada.

3.3. Mineração

Com o objetivo de testar diferentes algoritmos em diferentes situações, três novas bases foram criadas a partir de todo o tratamento já mencionado realizado na base de dados original e nas imagens disponibilizadas.

A primeira das bases, nomeada "*Base Original + 1 Atributo*", é a base originalmente disponibilizada devidamente tratada com um atributo adicional. Este atributo corresponde a uma discretização – em suma, a criação de faixas de valor e a atribuição dos dados a essas faixas - entre os valores de todas as seis informações retiradas de cada imagem. A discretização realizada tem como base os critérios adotados por Gonzalez e Woods (2009) que se encontram representados pela Tabela 1.

Tabela 1: Medidas de Textura. Gonzalez, Rafael C. Processamento digital de imagens.

Textura	Média	Desvio padrão	R normalizado	3º. momento	Uniformidade	Entropia
Suave	82,64	11,79	0,002	-0,151	0,026	5,434
Regular	99,72	33,73	0,017	-0,105	0,013	6,674
Rugosa	143,56	74,63	0,079	0,750	0,005	7,783

Portanto, um registro receberá, por exemplo, o valor 'Suave' para o seu atributo 'Textura' caso os atributos estejam nas seguintes faixas de valor: Média $\leq$ 82,64; Desvio Padrão $\leq$ 11,79; Descritor de Suavidade Relativa (R normalizado) $\leq$ 0,002; Terceiro Momento $\leq$ - 0,151; Uniformidade $\geq$ 0,026; Entropia $\leq$ 5,434.

A segunda base, nomeada "*Base Original + 4 Atributos*", trata-se da base originalmente disponibilizada devidamente tratada com os seguintes atributos adicionais: Descritor de Suavidade Relativa, Terceiro Momento, Uniformidade e Entropia. A terceira e última base, nomeada "*Base Original + 6 Atributos*", trata-se da base originalmente disponibilizada devidamente tratada com os seis atributos relatados a seguir: Média, Desvio Padrão, Descritor de Suavidade Relativa, Terceiro Momento, Uniformidade e Entropia. Com isso, executaremos os algoritmos em cada uma das bases e poderemos comparar as influências das diferenças apresentadas por cada uma delas nos resultados obtidos.

4. Resultados

4.1 Seleção do Algoritmo de Agrupamento

Para cada uma das bases os quatro algoritmos citados na seção 3.3.1 foram aplicados. Como forma de analisar os algoritmos de aprendizado não supervisionado as seguintes métricas foram utilizadas: *Coeficiente de Silhueta*, *Índice Calinski-Harabasz*, *índice*

Davies-Bouldin e *Soma dos Erros Quadráticos* [Pedregosa et al. 2011]. Todas foram utilizadas com seus parâmetros na forma padrão.

Para cada uma das métricas, a taxa de melhora média foi calculada através da média aritmética entre a taxa de melhora apresentada pela métrica em cada uma das três bases especificadas na seção 3.4. Para todos os algoritmos mencionados na seção 3.2.1 existe um parâmetro que especifica qual a quantidade de grupos a serem formados. Foram realizados testes com esse parâmetro assumindo os valores: 2, 4, 6 e 8. Pôde-se constatar que todos os algoritmos obtiveram um agrupamento melhor entre os valores 4 e 6; assim sendo, convencionou-se que a quantidade de grupos a serem formados pelos algoritmos em testes futuros seria 5. Após teste realizados, pôde-se constatar os seguintes resultados apresentados pelo gráfico 1.

Através da análise do gráfico é possível constatar que, para a métrica 'Soma dos Erros Quadráticos', o algoritmo KMeans se assemelha bastante ao MiniBachKMeans, destacando-se como o melhor através de uma diferença de 0,01%. É importante lembrar de que a métrica 'Soma dos Erros Quadráticos' não é compatível com os algoritmos Agglomerative e Birch Clustering; por tal motivo foi adotado que eles receberiam um valor padrão de 0. Na métrica 'Coeficiente de Silhueta' o algoritmo que se destaca é MiniBachKMeans sendo 32,26% superior ao segundo colocado, KMeans. O Calinski-Harabasz Index, por sua vez, apresenta o KMeans como sendo o algoritmo de melhor performance, diferenciando-se do segundo colocado, MiniBatchKMeans, por 8,65%.

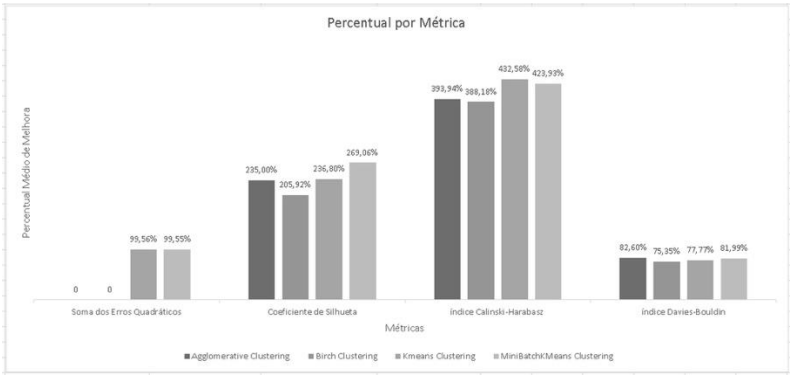

Gráfico 1: Percentual Médio de Melhora x Métricas

Por fim, para a métrica 'Índice Davies-Bouldin' temos o algoritmo Agglomerative como o de melhor performance, se destacando do segundo colocado, MiniBatchKMeans, por apenas 0,61%. Assim sendo, o KMeans Clustering foi selecionado para realizar o agrupamento dos dados.

4.2 Processo de Rotulação dos dados Agrupados

Uma vez definindo qual o melhor algoritmo de clusterização (KMeans), pode-se aplicar o algoritmo do rotulador [Lucas et al. 2014]. O retorno dado pelo rotulador está estruturado da seguinte forma: para cada grupo formado temos o seu identificador – que no caso, é um número – e os atributos que mais influenciam na caracterização desse grupo. Segundo Lucas et al. (2014) cada um desses atributos, juntamente com o seu valor mais recorrente, caracteriza-se como um rótulo A Tabela 2 contém os resultados apresentados pelo rotulador.

Tabela 2: Grupos formados pelo rotulador.

Grupo 0	
Atributos Rótulo	Valor
Matriz mesangial	Aumentada
Celularidade mesangial	Aumentada
Podocito	Normal
Necrose tubular aguda	Ausente
Vacuolização do epitelio	Ausente

Grupo 1	
Atributos Rótulo	Valor
Cristais	Ausente
Calcificação	Ausente
Podocito	Normal
Necrose tubular aguda	Ausente
Vacuolização do epitelio	Ausente

Grupo 3	
Atributos Rótulo	Valor
Cristais	Ausente
Calcificação	Ausente
Atrofia tubular	Ausente
Necrose tubular aguda	Ausente
Graduação de atrofia	Ausente

Grupo 2	
Atributos Rótulo	Valor
Cristais	Ausente
Tubulite	Ausente
Vasculite	Ausente
Necrose tubular aguda	Ausente
Vacuolização do epitelio	Ausente

Grupo 4	
Atributos Rótulo	Valor
Matriz mesangial	Aumentada
Celularidade mesangial	Aumentada
Cilindros	Hialinos
Imunofluorescência	Positiva
Tufo glomerular	Hipercelularidade endocapilar

Para cada um dos grupos formados é possível garantir que os atributos envolvidos são característica de todos dos registros que fazem parte dele, pois o valor mínimo de influência adotado no rotulador para o exemplo em questão foi de 100%. Em outras palavras, todos os elementos de um mesmo grupo possuem tais atributos em comum. Analisando a composição dos grupos é possível constatar que o atributo 'Necrose tubular aguda' está presente em quase todos os grupos – somente inexistente no grupo 4. Pode-se concluir que essa é uma característica em comum aos quatro outros grupos de acordo com a descrição resultante do algoritmo de clusterização. Tal interpretação é justificada pelo fato de todos os registros da base assumirem o valor 'ausente' para esse atributo.

Os atributos 'Podocito', 'Necrose tubular aguda' e 'Vacualização do epitelio' estão presentes nos grupos 0 e 1, o que demonstra uma possível semelhança entre os elementos desses grupos. Pode-se concluir, portanto, que os atributos 'Matriz mesangial' e 'Celularidade mesangial' – presentes no grupo 0 – 'Cristais' e 'Calcificação' – presentes no grupo 1 – demonstram um potencial diferencial de caracterização desses grupos.

O grupo 2 apresenta exclusivamente os atributos 'Tubulite' e 'Vasculite'. O grupo 3, por sua vez, é o único a apresentar os atributos 'Atrofia tubular' e 'Graduação de atrofia'. Ambos apresentam a menor quantidade de atributos rótulo dentre os cinco grupos formados. Tal realidade, juntamente com o fato de que a maioria dos seus atributos não são exclusivos do seu grupo, demonstra uma tendência dos grupos a não terem doenças claramente associadas a eles. Por fim, o grupo 4 apresenta com exclusividade os atributos 'Cilindros', 'Imunofluorescência' e 'Tufo glomerular'. Esses atributos demonstram um potencial diferencial entre o grupo 4 e os grupos restantes.

5. Conclusão

Com o fito de avaliar os grupos gerados pelo algoritmo e seus rótulos, foi solicitado aos mesmos profissionais da área da saúde que forneceram os dados iniciais (base de dados com patologias e as imagens renais) que analisassem os rótulos dos grupos formados e tentassem definir, levando em consideração os seus conhecimentos médicos, as doenças pertencentes a cada grupo. No Grupo 0, o especialista pode identificar *Nefropatia por IgA, Doença de Lesões mínimas, Podocitopatias e Nefrite lúpica classe II*. Já no Grupo 1 o especialista identificou que neste grupo pode conter diagnósticos de *Doença de membrana fina e/ou Doença de Lesões mínimas*. Assim como no grupo 4, de acordo com o rótulo foram identificados como possíveis diagnósticos *Glomerulonefrite membranoproliferativa, Nefrite lúpica classe III e Nefrite lúpica classe IV*. Infelizmente, nos grupos 2 e 3 nada pode ser concluído.

Dentre os cinco grupos formados, três tiveram doenças associadas a eles através da análise dos atributos rótulo apresentados. Pode-se concluir, portanto, que o algoritmo promove, dentro das limitações inerentes à base de dados, uma considerável qualidade de informação quanto ao agrupamento de registros semelhantes e quanto às suas respectivas características. Como podemos observar, é notável a influência positiva dos atributos retirados das imagens disponibilizadas quanto ao desempenho dos algoritmos de agrupamento na formação dos grupos. Além disso, constata-se a influência positiva do rotulador quanto ao processo de aquisição de conhecimento através das informações fornecidas por seu intermédio.

Assim, pode-se concluir que um bom processo de agrupamento, aliado ao processo de rotulação, promove uma significativa assistência aos profissionais da área quanto ao entendimento das características das patologias envolvidas e consequentemente auxilia no diagnóstico das doenças renais. Como propostas para trabalhos futuros, pode-se testar diferentes variações da quantidade de grupos a serem formados, assim como extrair novos descritores das imagens e implementar outros algoritmos de agrupamento.

6. Referências

Fayyad, U.M., Piatetsky-shapiro, G., Smyth, P. (1996) “Data Mining to Knowledge in Databases”, AI Magazine, Menlo Park, v.17.

Gonzalez, R.C., Woods, R.E. (2009) “Digital image processing”, Prentice Hall, Nova Jersey, v.2.

Santos, C.N. (2005) “Aprendizado de máquina na identificação de sintagmas nominais: O caso do português brasileiro”

Barber, D. (2012) “Bayesian Reasoning and Machine Learning”, Cambridge University, New York, NY.

Lucas, A.L., Vinicius, P.M. and Ricardo, A.L.R (2014) “Automatic Cluster Labeling through Artificial Neural Networks” International Joint Conference on Neural Networks (IJCNN), Beijin, China.

Pedregosa, et al. (2011) “Scikit-learn: Machine Learning in Python”, JMLR 12, pp. 2825-2830, https://scikit-learn.org/stable/modules/clustering.html#

Weiner, E. B.O., Alisson, F.P., Sandro, R.F., Silvana, T.F. (2014) “Classificação de Padrões Utilizando Descritores de Textura”, SIMMEC/EMMCOMP, Juiz de Fora, MG.