Diagnóstico das doenças eritemato escamosas com mineração de Dados.

1st Kaio Fabio Barbosa Rodrigues
Bacharelado em Sistemas de Informacão
Instituto Federal do Espirito Santo
Serra, Brazil
kaiofabiogs06@gmail.com

2nd Mateus Damaceno Schineider
Bacharelado em Sistemas de Informacão
Instituto Federal do Espirito Santo
Serra, Brazil
schiender.mateus@gmail.com

***Abstract*—As doenças Eritomato Escamosa são um grupo de 6 doenças que possuem características semelhantes em estágios distintos da doença, sua diagnosticação fica prejudicada e passível de erros devido a essas semelhanças. Por isso, este trabalho tem como objetivo submeter uma base de 366 Registros, contando com 34 Características de indivíduos distintos que possuem alguma das doenças Eritomato Escamosas, á algoritmos K-Means e Árvore de Decisão para realizar respectivamente agrupamento e classificação dos registros da base de dados.**

***Index Terms*—K-means, árvore de decisão, eritomato escamosa.**

I. Introdução.

Este trabalho tem o objetivo de avaliar a aplicação de algoritmos de agrupamento e classificação, no auxílio de doenças Eritemato Escamosas. Para isso, serão aplicados os algoritmos K-Means e Árvore de decisão em uma base de dados que conta com 366 registros e cerca de 34 Características. A base de dados tem suas características divididas em Características Clínicas(12 Atributos) e Características ”Histopatológicas”(22 Atributos).

Uma das motivações para aplicação dessa análise nos dados, é alta semelhança entre as características e sintomas das doenças presentes no grupo Eritemato Escamosas. Essa alta semelhança torna o diagnóstico das doenças passível de vários erros, sendo o diagnóstico influenciado por uma série de fatores externos aos sintomas como : Experiência do Responsável, Contexto dos Sintomas e etc.

Com o avanço da tecnologia, se vê a possibilidade do uso de soluções tecnológicas para auxiliar no diagnóstico e no aprendizado desse grupo de doenças, suas características e sintomas. Essas soluções serão representadas pelos métodos de agrupamento e classificação dos conceitos de Data Mining neste estudo.

II. Referencial Teórico

A. K-Means

O algoritmo K-Means é um algoritmo de aprendizado de máquina não supervisionado, baseado em agrupamentos [1].Os algoritmos de aprendizado não supervisionado são algoritmos que não recebem o resultado daquelas características, ou seja, recebem um aglomerado de informações e as agrupa de forma que os indivíduos com características semelhantes fiquem sempre próximos.

O K-Means utiliza o conceito de clusterização e centróides, “K” é o valor que se refere à quantidade de grupos a serem realizados pelo algoritmo, e centróide é basicamente o registro central daquele grupo de características. A partir das etapas do algoritmo, esse centróide é reposicionado e então reavaliado a posição dos demais registros dentro da base de dados.

É importante ressaltar que o K-Means pode emitir resultados diferentes em execuções diferentes na mesma base de dados e configurações, isso deve-se ao fato da sensibilidade do algoritmo em relação à sua escolha para centróide inicial, essa escolha interfere diretamente nas demais etapas do algoritmo [2].

A execução do algoritmo possui as seguintes etapas [1]:

- O primeiro passo na aplicação do algoritmo, é definir a quantidade de clusters(grupos), a serem montados pelo algoritmo, basicamente definir K.
- Após isso, o algoritmo define um centróide aleatório para cada cluster.
- Para cada registro da base de dados, é calculada sua distância em relação a cada centróide. O algoritmo escolhe o centróide mais próximo do ponto e o posiciona próximo a esse centróide.
- Para cada cluster, o algoritmo irá reposicionar seu centróide, baseado na média das posições de todos os registros agrupados nesse cluster.
- O algoritmo irá repetir todos os passos 2,3,4 até que a posição dos centróides de cada cluster se aproxime ao máximo da média dos registros daquele cluster.

A Figura 1 mostra a dispersão de dados da aplicação do algoritmo realizando agrupamento em 4 Clusters. Note que ao centro de cada cluster é possível identificar o centróide, sendo o registro que mais se aproxima da média dos registros pertencentes á aquele cluster.

B. Árvore de Decisão

Árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para classificar e prever dados. Em outras palavras, em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas.

Identify applicable funding agency here. If none, delete this.

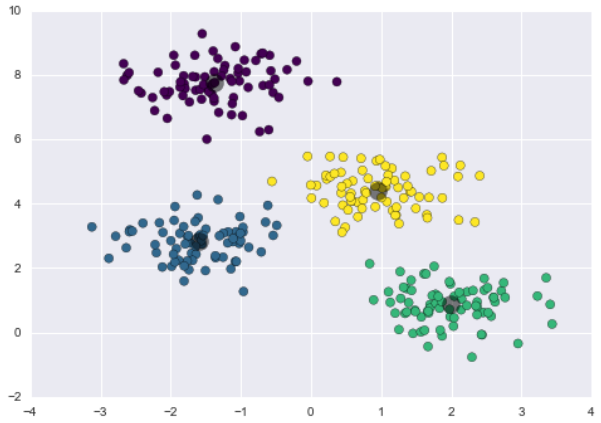

Fig. 1. Exemplo de Clusterização;

Estes modelos utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema [9].

Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore.

Existem três tipos de nós: nós de probabilidade, nós de decisão e nós de término. O nó de probabilidade, representado por um circulo, mostra as probabilidades de certos resultados. Um nó de decisão, representado por um quadrado, mostra uma decisão a ser tomada, e um nó de término mostra o resultado final de um caminho de decisão.

III. Metodologia

A base de dados utilizada para análise e aplicação dos algoritmos é nomeada de Dermatology Dataset e tem como principais desenvolvedores Nilsel Ilter e H. Altay Guvenir [**?**]. A base conta com 366 registros, com 34 características, sendo a 35ª o diagnóstico encontrado para o paciente com os sintomas descritos pelas 34 demais características.

As características são divididas entre características clínicas e características histopatológicas. As possíveis doenças diagnosticadas são divididas da seguinte forma e frequência:

TABLE I
Tabela de Clusters e Frequências

Código	***Doença***	***Diagnósticos***
1	Psoríase	112
2	Dermatite seborreica	61
3	Líquen plano	72
4	Pitiríase rósea	49
5	Dermatite crônica	52
6	Pitiríase rubra pilar	20

Para iniciar as análises e aplicação dos algoritmos, é necessário validar as informações e a consistência da base de dados. Para isso foi realizado o pré-processamento das informações removendo todo registro que não fosse inteiro da base de dados, além de remover todo registro que não estivesse preenchido ou com alguma informação faltante.

Após o pré-processamento, foi iniciado a etapa de aplicação do algoritmo K-means, como temos cerca de 6 possíveis grupos, a quantidade de clusters a serem construídos pelo algoritmo será 6. A aplicação deste algoritmo na base de dados foi realizada em 3 etapas. Cada etapa considerou uma determinada quantidade de características para realizar o treinamento e processamento do algoritmo.

Etapas realizadas na aplicação da análise da base utilizando o algoritmo K-means:

- A Primeira etapa consiste na aplicação do algoritmo em todos os atributos e características do registro, e então calcular sua precisão
- A Segunda etapa consiste na aplicação do algoritmo em todos os atributos e características clínicas do registro, e então calcular sua precisão.
- A Terceira etapa consiste na aplicação do algoritmo em todos os atributos e características histopatológicas do registro, e então calcular sua precisão.

O cálculo da precisão do algoritmo K-means foi feito baseando a coluna de diagnóstico do dataset com o agrupamento encontrado pelo algoritmo para o registro, portanto segue a seguinte fórmula $QtdAcertos/QtdAnalisado$.

Já para a aplicação da Árvore de Decisão, foi utilizado a lib $DecisionTreeClassifier$ presente na biblioteca $sklearn.tree$ do $Pythonversão3.8$. Durante a aplicação, para o treino do algoritmo, utilizou-se como critério o parametro "family history" presente na base de dados.

Durante a aplicação do algoritmo, foram dados os seguintes passos:

- Primeiro foi feito o treinamento do algoritmo, separando os dados de treino e os dados de teste passando como valores, 70% para treino e 30% para teste.
- A Segunda parte consiste na instanciação do objeto, para isso usou-se como critério de parada a métrica do grau de pureza dos dados, $entropia$.
- A Terceira etapa consiste na exibição da árvore resultante, para isso foi utilizado o $graphviz$.

IV. Resultados

Após a aplicação do algoritmo K-Means, foram obtidos os seguintes resultados em cada etapa.

TABLE II
Tabela de Clusters e Frequências

Etapa	***Precisão***
1. (Todos Atributos)	0.1284
2. (Atributos Clínicos)	0.2076
3. (Atributos Histopatológicos)	0.2759

É possível perceber que ao aplicar o algoritmo a todos os atributos da base de dados, teve-se uma redução considerável de 38-53% se comparado com as etapas 2 e etapa 3 que levaram em consideração somente os atributos específicos. E

```
from sklearn import metrics


print (metrics.classification_report(y_test, resultado))
```

	precision	recall	f1-score	support
0	0.89	0.91	0.90	96
1	0.25	0.21	0.23	14
accuracy			0.82	110
macro avg	0.57	0.56	0.56	110
weighted avg	0.81	0.82	0.81	110

Fig. 2. Resultado Árvore de Decisão;

que ao analisar somente as características Histopatológicas, a precisão do algoritmo foi de até 53% maior que as demais etapas. No geral, todas as etapas tiveram uma precisão bem abaixo do esperado, visto que menos que 30% de precisão não demonstra capacidade de realizar o diagnóstico dos casos de forma independente.

Após a aplicação da Árvore de decisão, obtivemos os resultados visiveis na Figura 2.

Como podemos perceber, a precisão do algoritmo foi de 82%. Utilizando como critério de parada a métrica de impureza gini, não foram tão satisfatórios os resultados quanto os utilizando entropia.

V. Conclusão

Após análise dos resultados do algoritmo K-Means após aplicados na base dados, foi verificado que sua precisão está bem abaixo do esperado, visto que o mínimo esperado para atender um diagnóstico automático seria de no mínimo 80%, visto a importância do processo diagnosticação das doenças. Portanto, para este trabalho, o K-Means se tornou ineficiente para realizar o agrupamento dos registros nos clusters esperados.

Já com precisão em torno dos 82% os resultados obtidos com a aplicação da Árvore de Decisão se mostraram satisfatórios, atendendo as expectativas para a analise do processo de diagnosticação das doenças.

References

[1] B. Anastacio. Programadores Ajudando Programadores.
[2] P. Sampaio. Entendendo K-Means.
[3] Nilsel Ilter e H. Altay Guvenir.
[4] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[5] K. Elissa, “Title of paper if known,” unpublished.
[6] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[8] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
[9] J. Gama, ”Functional Trees”, Machine Learning, 55, 219–250, 2004, Kluwer Academic Publishers.