Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Rotulação com Algoritmos Supervisionados

Tarcísio Franco Jaime

Número de Ordem PPGCC: M001

Teresina-PI, Janeiro de 2017

Tarcísio Franco Jaime

Rotulação com Algoritmos Supervisionados

Qualificação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Vinicius Ponte Machado

Teresina-PI

Janeiro de 2017

Tarcísio Franco Jaime
Rotulação com Algoritmos Supervisionados/ Tarcísio Franco Jaime. – Teresina-PI, Janeiro de 2017-
43 p. : il. (algumas color.) ; 30 cm.

Orientador: Vinicius Ponte Machado

Qualificação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, Janeiro de 2017.

1. Rotulação. 2. Algoritmos Supervisionados. 3. CART. 4. Naive Bayes. I. Vinicius Ponte Machado. II. Universidade Federal do Piauí. III. Rotulação com Algoritmos Supervisionados.

CDU 02:141:005.7

Tarcísio Franco Jaime

Rotulação com Algoritmos Supervisionados

Qualificação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 01 de janeiro de 2018:

Vinicius Ponte Machado
Orientador

Co-Orientador

Professor
Convidado 1

Professor
Convidado 2

Professor
Convidado 3

Teresina-PI
Janeiro de 2017

Aos meus pais XXXXXXXX e YYYYYYY,
por sempre estarem comigo em todos os momentos.

Agradecimentos

Agradeço a Deus.

Agradeço aos meus pais, XXXXX e YYYYY, por ...

Aos meus irmãos, por.....

Agradeço ao meu orientador, XXXXXXXXX, por todos os conselhos, pela paciência e ajuda nesse período.

Aos meus amigos ...

Aos professores ...

À XXXXXX pelo apoio financeiro para realização deste trabalho de pesquisa.

"Não sei o que,
não sei o que,
não sei o que lá."
(Autor Desconhecido)

Resumo

Segundo a ABNT, o resumo deve ressaltar o objetivo, o método, os resultados e as conclusões do documento. A ordem e a extensão destes itens dependem do tipo de resumo (informativo ou indicativo) e do tratamento que cada item recebe no documento original. O resumo deve ser precedido da referência do documento, com exceção do resumo inserido no próprio documento. (. . .) As palavras-chave devem figurar logo abaixo do resumo, antecedidas da expressão Palavras-chave:, separadas entre si por ponto e finalizadas também por ponto.

Palavras-chaves: latex. abntex. editoração de texto.

Abstract

This is the english abstract.

Keywords: latex. abntex. text editoration.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

EWD	Discretização por Larguras Iguais
EFD	Discretização por Frequências Iguais

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

Sumário

Introdução

Este documento segue as normas estabelecidas pela **??**, 3.1-3.2).

A proposta deste mestrado bem como outros trabalhos relacionados, onde áreas envolvidas tem como tema principal, Rotulação de Dados, estão alterando a maneira de como Aprendizagem de Máquina define este termo. Em pesquisas realizadas neste área sob supervisão do orientador desta proposta, vários trabalhos estão definindo Rotulação sendo algo diferente da Classificação dos dados.

Apesar de várias literaturas (BARBER, 2011; MITCHELL, 1997) entre outras citarem o termo rotulação como um sinônimo de classificação, neste departamento, esse termo esta ficando obsoleto. Muitos trabalhos feitos aqui neste laboratório estão redefinindo o termo rotulação como algo mais completo e que possui propriedade diferente a apresentada na classificação.

A classificação é dada com um identificador do registro informanto qual classe ele pertence....

Figuras

As normas da **??**, 3.1-3.2) especificam que o caption da figura deve vir abaixo da mesma.

A Figura 1 ilustra...

Figura 1 – Breve explicação sobre a figura. Deve vir abaixo da mesma.

Tabelas

A Tabela 1 apresenta os resultados...

Tabela 1 – Breve explicação sobre a tabela. Deve vir acima da mesma.

XX	FF	PP	YY	Yr	xY	Yx	ZZ
615	18	2558	0,9930	0,9930	0,9930	0,9930	0,9930
615	18	2558	0,9930	0,9930	0,9930	0,9930	0,9930
615	18	2558	0,9930	0,9930	0,9930	0,9930	0,9930
615	18	2558	0,9930	0,9930	0,9930	0,9930	0,9930
615	18	2558	0,9930	0,9930	0,9930	0,9930	0,9930

Motivação

Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

Objetivos

Nulla malesuada risus ut urna. Aenean pretium velit sit amet metus. Duis iaculis. In hac habitasse platea dictumst. Nullam molestie turpis eget nisl. Duis a massa id pede dapibus ultricies. Sed eu leo. In at mauris sit amet tortor bibendum varius. Phasellus justo risus, posuere in, sagittis ac, varius vel, tortor. Quisque id enim. Phasellus consequat, libero pretium nonummy fringilla, tortor lacus vestibulum nunc, ut rhoncus ligula neque id justo. Nullam accumsan euismod nunc. Proin vitae ipsum ac metus dictum tempus. Nam ut wisi. Quisque tortor felis, interdum ac, sodales a, semper a, sem. Curabitur in velit sit amet dui tristique sodales. Vivamus mauris pede, lacinia eget, pellentesque quis, scelerisque eu, est. Aliquam risus. Quisque bibendum pede eu dolor.

1 Referencial Teórico

Será abordado neste capítulo o conteúdo base na compreensão deste trabalho dividido em 3 sessões: Aprendizado de Máquina, Discretização e Trabalhos Correlatos.

A primeira sessão contempla os principais tipos de aprendizados indutivos, não incluindo aqui o aprendizado semi-supervisionado e sim dando ênfase a aprendizagem supervisionada, foco da proposta deste mestrado. O aprendizado indutivo utiliza uma amostra do todo para tirar uma conclusão. Caso os exemplos retirados de uma base de dados não forem suficientes, talvez o conhecimento derivado destes exemplos não mostrem a verdade.

O segundo ítem dissertará sobre a técnica de discretização adotada nesta pesquisa. Possuindo grande contribuição para os resultados gerados, e ganhando assim uma sessão própria para explanação de como funciona essa técnica. E na terceira sessão serão abordados trabalhos com mesmas características particulares para melhor elucidar o motivo da elaboração dessa proposta de mestrado.

1.1 Aprendizado de Máquina

Aprendizagem de máquina é a capacidade do aprendizado automático com utilização de algoritmos atuando em cima de uma base de dados. Diz-se que o computador está aprendendo quando existe uma melhora de desempenho de tarefas que ele utilizou como exemplo (MITCHELL, 1997). Um exemplo seria a realização do reconhecimento facial de uma pessoa utilizando aprendizado de máquina. Não seria necessário a implementação de várias linhas de código informando que a cor dos olhos são azuis com orelhas e cabelos grandes, seriam de uma certa pessoa. Ao invés disso é observada várias fotos tituladas de uma certa pessoa, e após vários exemplos o computador seria capaz de predizer uma foto nova, se é, ou não, da determinada pessoa através de aprendizado anterior.

Existem alguns motivos, onde justificam, que não é possível simplismente exigir que o projetista implemente melhorias no sistema de forma que ele esteja robusto bastante para lidar com todas as situações (RUSSEL; NORVIG, 2013). Um desses motivos seria a incapacidade da antecipação de todas as situações possíveis de implementação por parte do programador. Fazendo um resumo, aprendizado de máquina seriam algoritmos capazes de aprender automaticamente através de determinados exemplos, ou comportamentos.

A partir desta síntese, tem-se uma observação. A classificação de dados no contexto de aprendizado de máquina, são compostos por dois pilares. Um, seriam os **dados** a serem classificados, e outro, o **algoritmo** que irá atuar nessa base de dados. Existem vários

algoritmos como exemplo: redes neurais, árvores de decisão, Suport Vector Machine – SVM, etc. Qualquer um destes algoritmos são utilizados para solucionar essa classificação. E a escolha apropriada, desse algoritmo, se dará através de métricas que avaliarão o desempenho de cada um, e a melhor métrica, será o algoritmo apropriado para aquele problema de classificação de dados.

Uma analogia referente do que foi dito acima seria um “problema”, comparado a um “motor”, e os algoritmos disponíveis seriam as "ferramentas"para concertar esse motor. A partir daí a ferramenta que fosse mais eficaz, considerando métricas de desempenho, para fazer o motor funcionar, seria a ferramenta(algoritmo) escolhida. Tendo assim a escolha certa para um determinado problema.

1.1.1 Aprendizado Supervisionado

Nesta sessão será abordado um método que através de uma banco de dados já classificado por especialistas, será feita uma predição de novos registros com base em vários desses exemplos já classificados. Os responsáveis por essas predições de novos registros são algoritmos de aprendizado supervisionados projetados para determinados fins.

O termo "Supervisionado"indica que existe um supervisor para cada registro de entrada especificando uma saída para esse registro. Considerando uma base de dados de imagens de rostos, onde cada imagen possui uma saída representado por uma classe: masculino ou feminino. A tarefa seria criar um preditor capaz de acertar a cada novo registro se a imagem é masculina ou feminina. Seria difícil implementar de maneira tradicional, uma vez que são inúmeras as diferenças que difere as faces masculinas e femininas. Mas uma alternativa seria dar exemplos de rostos com suas classificações de fazer que automaticamente a máquina "aprenda"uma regra para predizer se é masculino ou feminino (BARBER, 2011).

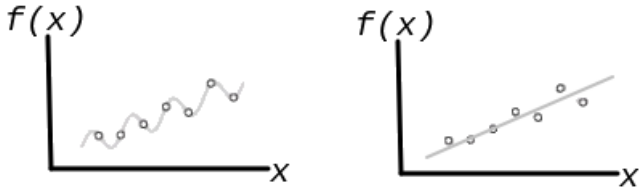
Em (RUSSEL; NORVIG, 2013) os autores fazem uma apresentação formal do funcionamento da aprendizagem supervisionada. Dado um conjunto de treinamento

$$(x_1, y_2), (x_2, y_2), ...(x_n, y_n), \tag{1.1}$$

onde cada y_j foi gerado por $y = f(x)$ desconhecida. Encontrar uma função h que se aproxime da função f real.

A função h é uma hipótese onde prevê um melhor desempenho entre as hipóteses possíveis através dos conjuntos de exemplos, que são diferentes do conjunto de treinamento 1.1.

Na figura 2a existe um sobre ajuste da função com o conjunto de dados de treinamento. Esse exemplo acabou exibindo uma função mais complexa para se molda de acordo com os sete pontos do gráfico, especificando para esse conjunto de dados.



(a) Ajuste polinomial de grau 6

(b) Hipótese linear

Figura 2 – Hipóteses ajustadas

Ja na figura 2b o ajuste da função se torna mais simples e mesmo não passando por todos os pontos, acabou generalizando melhor o conjunto de treinamento, tornando talvez, um melhor resultado da predição de novos valores.

A figura 2 mostra duas hipóteses que tentam se aproximar ao máximo da função verdadeira, que é desconhecida. Mesmo parecendo que na figura 2a obteve-se melhor resultado, pois todos os pontos são contemplados pela função, mas esta função h acabou ficando muito específica e isso não retrata os dados em um mundo real. Então quanto mais generalizado for h, melhor será para prevê os valores de y para novos conjuntos de dados.

Antes de falar dos algoritmos utilizados nesse texto a aprendizagem supervisionada detem dois tipos de caso: regressão e classificação. A classificação, contêm variáveis com valores discretos, onde as amostras destas variáveis de saída estão na forma de categorias. Como exemplo poderia ser masculino e feminino. Já no tipo regressão, possuem valores contínuos: quantidade de água em ml, velocidade de um carro, altura de uma pessoa.

1.1.1.1 Algoritmo Classification and Regression Trees - CART

Esse algoritmo constroi modelos de previsão a partir de dados de treinamento onde seus resultados podem ser reprensentados em uma árvore de decisão. No caso de não ser probabilístico o grau de confiança em seu modelo de predição será embasada em respostas semelhantes em outras circunstâncias antes analisadas.

Inicialmente todas as amostras se concentram no nó raiz, e a partir daí é apresentado uma questão, onde a intenção é separar o nó raiz em dois grupos mais homeogêneos. Dependendo da questão as amostras iram para a folha esqueda ou direita do nó raiz.

O CART faz essa divisão em função da regra Gini[1]**??**, parecida com a regra da entropia usada no algoritmo ID3[2]. O índice Gini varia de 0 a 1, definindo o grau de pureza do nó.

$$Gini(S) = 1 - \sum p^2(j/t) \tag{1.2}$$

[1] O CART pode utilizar outros critérios de divisão de dados como: entropia e critério de Twoing

[2] Algoritmo abordado por (**??**)

Onde: $p(j/t)$ é probabilidade a priori da classe j se formar no nó t. E S é um conjunto de dados que contém exemplos de n classes

Para construção de uma árvore existem três componente importantes (YOHANNES; WEBB, 1999):

- Um conjunto de perguntas que servirá de base para fazer uma divisão;
- Regras de divisão para julgar o quanto é boa esta divisão;
- Regras para atribuir uma classe a cada nó;

Abaixo segue um algoritmo de como o critério Gini é aplicado nas variáveis (RAIMUNDO; MATTOS; WALESKA, 2008):

Algorithm 1: Rotina de funcionamento do CART

```
1  melhorGini; /* cria a variável                                          */
2  divisaoCorrente ← 4.9;/* Ex. recebe o 1° valor do atributo              */
3  direita ← 0;
4  esquerda ← 6;/* Ex. recebe o total de dados existentes para o
       atributo                                                           */
5  while existirem dados do
6      if 1ª Dado Lista do Atributo MAIOR divisaoCorrente then
7          valorGini ← calculaGini(divisaoCorrente);
8      else
9          valorGini ← calculaGini(1ªDadoLista);
10     if Primeiro Gini encontrado then
11         melhorGini ← valorGini;
12     else
13         if valorGini > melhorGini then
14             melhorGini ← valorGini
15     divisaoCorrente ← 5.4; /* recebe o próximo dado do atributo          */
16     direita recebe o que possui +1 e esqueda o -1;
17     (valorGini + divisaoCorrente)/2;/* encontrar ponto de divisão       */
```

1.1.1.2 Algoritmo Naive Bayes

É um algoritmo considerado rápido, em relação a outros algoritmos de classificação, mesmo com grandes volumes de dados em seu conjunto de treinamentos. Utiliza modelo probabilístico, Teorema de Bayes e possue a característica de independência dos atributos, onde as classes não dependem de recursos de outras. Essa independência condicionada entre os atributos, os quais nem sempre ocorrem nos problemas reais, acabou sendo conhecida por Bayes ingênuo, ou Naive Bayes.

Naive Bayes como classificador estatístico possue um modelo de simples construção, e ficou conhecido por ter bons resultados em relação a algoritmos mais sofisticados, mesmo trabalhando com grandes quantidades de dados. Ele agrupa objetos de uma certa classe em razão da probabilidade do objeto pertencer a esta classe.

$$P(c/x) = \frac{P(x/c)P(c)}{P(x)} \tag{1.3}$$

$$P(c/x) = P(x_1|c) * P(x_2|c) * ... * P(x_n|c) * P(c) \tag{1.4}$$

- $P(c/x)$ probabilidade posterior da classe $c, alvo$ dada preditor $x, atributos$.
- $P(c)$ é a probabilidade original da classe.
- $P(x|c)$ é a probabilidade que representa a probabilidade de preditor dada a classe.
- $P(x)$ é a probabilidade original do preditor.

A utilização do algoritmo Naive Bayes já é bem difundida, e está presente em vários trabalhos, como classificação de textos, filtro de SPAM, analisador de sentimentos, entre outros (MADUREIRA, 2017; LUCCA et al., 2013; WU et al., 2008; MCCALLUM; NIGAM, 1997). Mas mesmo atingido popularidade existem pontos negativos. A suposição de ter preditores independentes não acontece muito na vida real, pois acaba sendo difícil ter uma amostra de dados que sejam inteiramentes independentes.

Outra situação é caso de existir uma variável categórica que não foi observada na amostra tirada para o conjunto de treinamento, então poderá o modelo atribuir probabilidade 0(zero), não sendo capaz de fazer uma previsão. Quando isso acontecer uma técnica de alisamento é aplicada, chamada estimativa de Laplace, utilizadas em probabilidades condicionadas.

1.1.2 Aprendizado Não Supervisionado

No Aprendizado Não Supervisionado , não existe uma tentativa de se encontrar uma função que se aproxime da real. Logo porque os registros não são classificados, então o conjunto de treinamento não possue informação da saída sobre determinada entrada . Desta forma os algoritmos procuram algum grau de similaridade entre os registros e tenta agrupá-los de forma a ter algum sentido deles estarem juntos.

Quando o algoritmo encontram dados com mesma similaridade ele os agrupa formando clusters. Os números de clusters encontrados irão depender de como os algoritmos funcionam, junto com o grau de dissimilaridade entre elementos de grupos diferentes. Como não existe uma variável classe no Aprendizado Não Supervisionado, então (BARBER,

2011) diz que o maior interesse seria em uma perspectiva probabilística de ditribuição $p(x)$ de um determinado conjunto de dados.

$$D = \{x_n, n = 1, ..., N\} \tag{1.5}$$

Uma vez que no conjunto 1.5 não existe classe y, encontrado em um conjunto de treinamento 1.1 o algoritmo precisa encontrar padrões nos atributos para fazer os agrupamentos.

1.2 Discretização

A discretização faz parte em duas etapas no modelo defendido nesse trabalho, por isso a preocupação na explanação de seu funcionamento aqui nesta sessão. O método de discretização faz a conversão de valores contínuos em valores discretos. A partir de um atributo com valores contínuos, a discretização irá forçar um ponto inicial e final definindo um intervalo e designando uma faixa para cada intervalo. Assim, ao invés de valores contínuos em cada atributo, será relacionado a faixa que aquele atributo pertence, definindo assim seu novo valor. O melhor método de discretização seria encontrar o conjunto de valores contínuos por faixa de intervalos pequenos (KOTSIANTIS; KANELLOPOULOS, 2006)

A partir de alguns autores (CATLETT, 2006; HWANG; LI, 2002) a discretização melhora a precisão e deixa um modelo classificador mais rápido em seu conjunto de treinamento. Aqui nesse trabalho é utilizado a técnica de discretização antes da execução dos algoritmos e as faixas selecionadas são usadas para identificar o rótulo. Após o conhecimento do rótulo o valor da faixa é trocado pelo início e fim do intervalo.

Os métodos de discretização mais comumente utilizados no âmbito dos métodos não-supervisionados de acordo com (KOTSIANTIS; KANELLOPOULOS, 2006; DOUGHERTY; KOHAVI; SAHAMI, 1995) são os métodos de Discretização por Larguras Iguais(EWD) e Discretização por Frequências Iguais (EFD).

1.2.1 Discretização por Larguras Iguais - EWD

O método de Discretização por Larguras Iguais (EWD) faz a discretização de um intervalo, entre valores contínuos, dividindo em faixas de tamanhos iguais. Logo se existir um intervalo com valores contínuos [a,b], e deseja particionar em R faixas de tamanhos iguais serão necessários $R - 1$ pontos de corte figura 3.

Para haver o ponto de corte antes tem que ser realizado a ordenação dos dados. A largura de cada faixa $r_1, ..., r_R$ na equação 1.6 é representada por w que é calculada pela

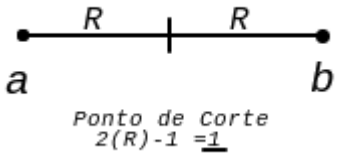


Figura 3 – Ponto de Corte (R-1)

diferença entre os limites superior e inferior do intervalo, dividido pela quantidade R de valores a serem gerados.

$$w = \frac{b-a}{R} \tag{1.6}$$

A variável w determina os pontos de corte $(c_1, ..., c_{R-1})$ que irão delimitar o tamanho das faixas de valores. O primeiro ponto de corte, c_1, é obtido através da soma do limite inferior a com a tamanho de w. E os pontos de corte seguintes são calculados pela soma do ponto de corte anterior com w.

O valor de cada faixa será representado por i, onde i é o índice indicando a faixa. De acordo com a figura 4 para dividir o intervalo $[a, b]$ em R faixas será necessário de $R-1$ pontos de corte.

$$c_i = \begin{cases} a+w, & se\, i = 1 \\ c_{i-1}+w, & caso\, contrário \end{cases} \tag{1.7}$$

O valor da faixa do intervalo $[a, c_1]$ será o valor discreto igual ao índice de sua faixa r_1. Então, um valor na faixa r_1 terá o valor reprensentado por $1(um)$, pois $i = 1$ é limite inferior mais largura da faixa, equação 1.7. E seguindo o mesmo raciocínio o valor da faixa $r_2 =]c_1, c_2]$ é reprensentado por $2(dois)$, e consequentemente o valor que se encontra em uma faixa qualquer r_i será reprensentado por i.

1.2.2 Discretização por Frequência Iguais - EFD

Esse outro método de discretização já possue uma abordagem diferente a do EWD, pois a idéia é manter a quantidade de elementos distintos, entre os pontos de corte, com o mesmo número. Dado um intervalo $[a, b]$ o número de faixas R e a quantidade de valores distintos ξ , onde $\xi \geqslant R$ o método EFD irá segmentar em R faixas de valores que possuem a mesma quantidade de elementos distintos λ. Então serão realizados $R-1$ pontos de corte gerando R faixas de valores, $(r_1, ..., r_R)$, com a mesma quantidade de elementos distintos λ. Para encontrar λ calcula-se o valor inteiro da divisão entre a quantidade de elementos

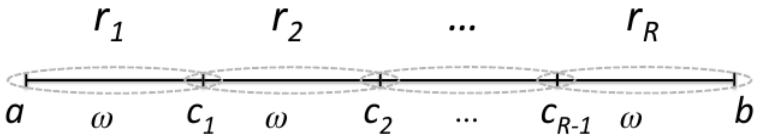


Figura 4 – Discretização EWD [3]

distintos ξ pela quantidade de faixas de valores R, obtendo o número de elementos da faixa 1.8.

$$\lambda = \frac{\xi}{R} \tag{1.8}$$

Uma observação nesse método é quando ocorrer nos casos de uma amostragem possuir uma má distribuição de valores de um dado atributo, como um número significativo de repetições, isso, irá causar um desiquilíbrio nas distribuições dos elementos.

Uma vez no intervalo $[a, b]$ de elemetos ordenado e calculado λ contendo R elementos ($v_{[R]}$ pode-se determinar os pontos de corte $(c_1, ..., c_{R-1})$ que são os delimitadores das faixas. Cada ponto de corte c_i pode ser calculado por $v_{i\lambda}$ 1.9.

$$\lambda = \frac{\xi}{R} \tag{1.9}$$

Como na sessão anterior do método EWD o valor que estiver no intervalo $[a, c_1]$ terá seu valor associado a um valor discreto igual ao índice i de sua faixa r_i conforme figura 5. Então, caso o valor esteja na faixa r_2 ele passará a ter o valor de seu índice i igual a $2(dois)$. De maneira consecutiva os valores que estiverem na faixa $r_3 =]c_2, c_3]$ terão valor $3(três)$. Uma outra observação desse método é que diferente do EWD, as faixas podem assumir faixas com tamanhos diferentes.

1.3 Trabalhos Correlatos

Esta sessão propõe relacionar outros trabalhos servindo de complemento teórico, como também leitura imprescindível, para entender a variedade de aplicações referente ao assunto de rotulação de dados. Mas ao longo da escrita desta proposta de mestrado verificou-se uma carência de pesquisas no âmbito de rotulação de dados, referente ao tema aqui proposto neste trabalho, pois acaba sendo redefinido o termo de rotulação.

O trabalho escrito por (LOPES; MACHADO; RABELO,) fez um estudo abordando o tema de rotulação de dados bastante significativo. Foi aprensentado nesse trabalho o Problema de Rotulação, que representa também o problema proposto por esse trabalho, mas com abrangência e execução diferente do modelo (LOPES; MACHADO; RABELO,)

[3] Figura extraída de (LOPES; MACHADO; RABELO,)

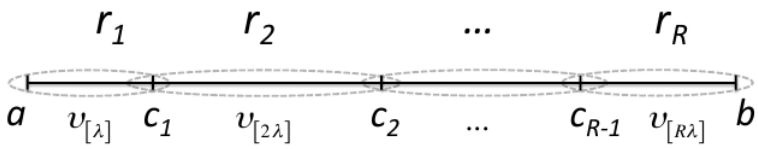


Figura 5 – Discretização EFD[4]

na figura 6 . Na pesquisa de (LOPES; MACHADO; RABELO,) é utilizado como entrada um conjunto dados onde é feito um agrupamento automático formando os clusters, e aprensenta como saída um rótulo específico que melhor define o grupo formado. Esses rótulos são formados pela faixa de valor em conjunto com os atributos mais relevantes.

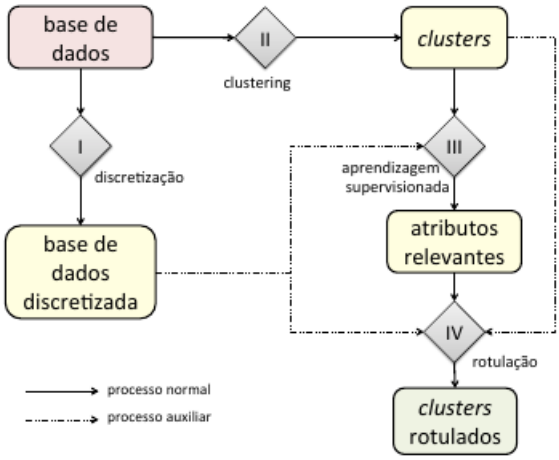


Figura 6 – Modelo (LOPES; MACHADO; RABELO,)

Outra pesquisa aplicada em rotulação está em (FILHO, 2015) onde aborda o mesmo Problema de Rotulação. Mas a atuação é diferenciada, pois o modelo, figura 7 procura diferenças existentes em cada grupo através da seleção dos elementos que representam o grupo, e depois é construído a faixa de valores. Os grupos são formados pelo algoritmo Fuzzy C-Means e após isso que é selecionado os atributos.

Em (LIMA, 2015) o problema em questão é fazer classificação e rotulação em uma base que possuem poucos elementos classificados. O método inicia com uma base dividida em elementos classificados(L) e não classificados(U). Após cada iteração o grupo L vai crescendo e automaticamente diminuindo o grupo U até que não tenha mais nenhum elemento em U. Após isso é realizado uma etapa de agrupamento, sem levar em consideração os dados classificados anteriormente. Terminada essa etapa é feito uma validação para saber quais os rótulos foram considerados corretos.

[4] Figura extraída de (LOPES; MACHADO; RABELO,)

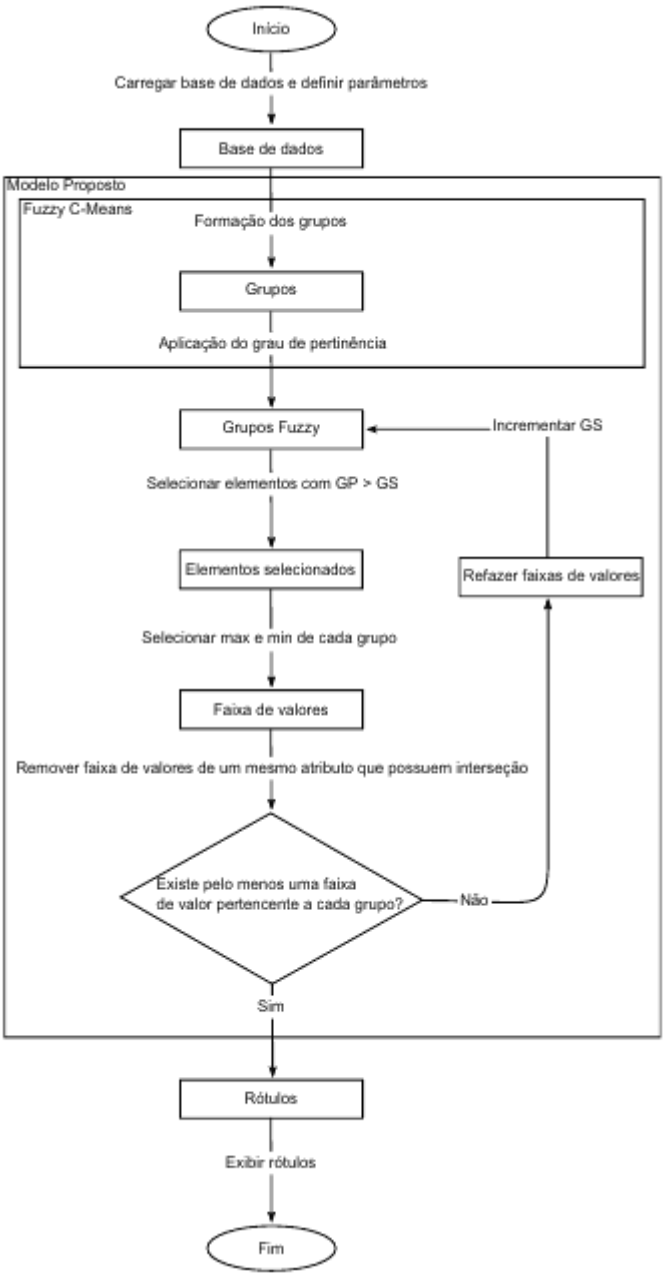


Figura 7 – Modelo (FILHO, 2015)

2 Metodologia / Materiais e Métodos

Esse capítulo abordará em uma sessão o problema proposto por esse trabalho, e logo em seguida, será apresentado um modelo de resolução. O objetivo ao final deste capítulo é poder resolucionar o problema, exibindo seus passos, e atribuir a qualquer outro pesquisador todo o conhecimento necessário para replicar este trabalho através das informações produzidas aqui.

2.1 Considerações do Problema

A abordagem do problema referente a essa proposta de mestrado segue uma linha já pequisada por (LOPES; MACHADO; RABELO,), que seria o **Problema de Rotulação**. Esse conceito, rotulação de dados, já é estudado na literatura na área de aprendizagem não-supervisionada, sessão 1.1.2, onde é comum os algoritmos lidarem com os agrupamentos dos dados, onde clusters são criados a partir dos graus similaridade entre os elementos.

Muitas pesquisas realizadas na área de rotulação fazem referencia, de fato, a classificação do dados, e não da rotulação, nos termos desse trabalho. Ao agrupar um conjunto de elementos por um derterminado critério, esta havendo uma classificação desses elementos escolhidos, mas pouco se sabe, qual é a compreensão desses grupos, já classificados.

Existe uma importância na criação dos clusters, contudo para o espectador é interessante existir um rótulo, desse grupo formado, oferecendo elementos em alguma tomada de decisão em razão de seu significado(rótulo).

Tem-se então o real problema de rotulação, contudo é necessário existir algum elemento definindo o porquê daquele grupo formado. O elemento é um rótulo composto por um, ou vários, atributo(s) de maior relevância no cluster, junto com uma faixa de valores. Essa faixa é um intervalo de valores definido pela discretização **??**, onde o intervalo escolhido, seria a faixa que apresenta os valores que se repetem com a maior frequência.

O Problema de Rotulação é formalmente definido como segue abaixo:

Definição 1 *Dado um conjunto de clusters* $C = \{c_1, ..., c_k | K \geqslant 1\}$*, de modo que cada cluster contém um conjunto de elementos* $c_i = \{\vec{e}_1, .., \vec{e}_{n^{(c_i)}} | n^{(c_i)} \geqslant 1\}$ *que podem ser representados por um vetor de atributos definidos em* $\mathbb{R}^m$ *e expresso por* $\vec{e}^{c_i} = (a_1, .., a_m)$ *e ainda que com* $c_i \cap c_{i'} = \{0\}$ *com* $1 \leqslant i, i \leqslant K$ *e* $i \neq i'$*.* [1]

[1] Adaptada de (LOPES; MACHADO; RABELO,)

- K *é o número de clusters;*
- c_i *é o i-ésimo cluster qualquer;*
- n^{c_i} *é o número de elementos do cluster* c_i*;*
- $\vec{e}_{n^{(c_i)}}$ *se refere ao j-ésimo elemento pertencente ao cluster* c_i*;*
- m *é a dimensão do problema;*

2.2 O Modelo de Resolução

Uma vez já conhecido a definição do problema - *Definição 1* - é possível situar a abrangência abordada aqui nessa pesquisa, pois a intenção do estudo científico desenvolvido aqui é provar a realização de **rotulação de dados com qualquer algoritmo supervisionado**, utilizando as técnicas abordadas neste texto.

O Modelo aqui proposto consiste em apresentar como saída um conjunto de rótulos, onde cada rótulo específico é dado por um conjunto de pares de valores, atributo e seus respectivo intevalor, gerado a partir da frequência de valores repetidos neste intervalo. Segue *Definição 2* formalizando a saída do modelo:

Definição 2 *Dado um conjunto de rótulos* $R = \{r_{c1}, ..., r_{ck}\}$*, no qual cada rótulo específico é dados por um conjunto de pares de valores, tem como saída um vetor com atributo e seu respectivo intervalo,* $r_{ci} = \{(a_1, [p_1, q_1]), ..., (a_{m^{(c_i)}},]p_{m^{(c_i)}}, q_{m^{(c_i)}}])\}$ *capaz de melhor expressar o cluster* c_i*.* [2]

- k *número de rótulos;*
- R *representa o conjunto de rótulos na saída do modelo;*
- a *é o atributo*
- c_i *é o i-ésimo cluster;*
- r_{c_i} *é o rótulo referente ao cluster* c_i*;*
- $]p_{m^{(c_i)}}, q_{m^{(c_i)}}]$ *representa o intervalo de valores do atributo* $a_{m^{(c_i)}}$*, onde* $p_{m^{(c_i)}}$ *é o limite inferior e* $q_{m^{(c_i)}}$ *é o limite superior;*
- m *é a dimensão do problema;*

Como apresentado na sessão 1.3, o autor foca em rotulação automática de grupos utilizando a estratégia de aprendizagem de máquina supervisionada, e paradigma conexionista, para provar seu trabalho. Mas aqui nessa pesquisa foi aplicado no modelo um acréscimo de 2(*dois*) algoritmos com paradigmas de aprendizado diferentes do que já foi utilizado, compondo uma base para afirmar, que a partir dessas amostras pode-se fazer rotulação com qualquer algoritmo supervisionada.

[2] Adaptada de (LOPES; MACHADO; RABELO,)

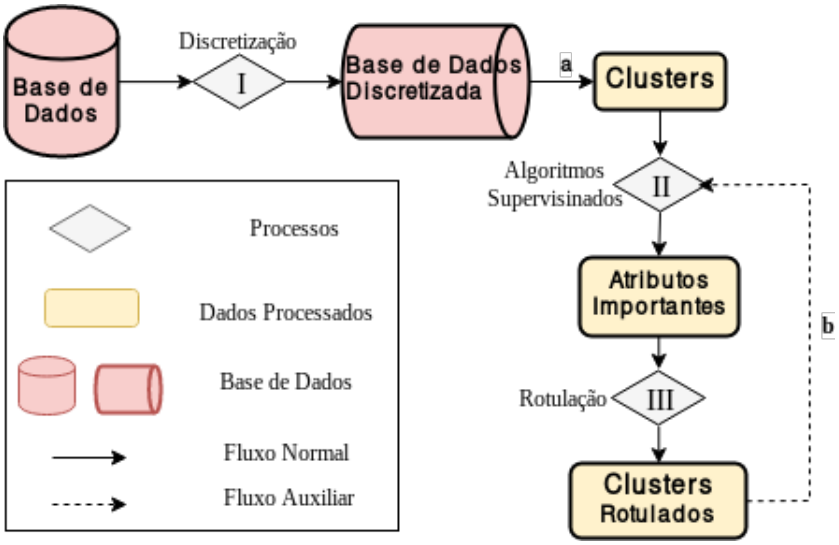


Figura 8 – Modelo de Resolução Proposto

O modelo, figura 8, inicialmente mostra a Base de Dados já classificada, pois o cerne desta pesquisa é conceder ao grupo um significado,a rotulação, através da técnica de correlação entre atributos, sessão 2.3. Essa base poderá conter valores contínuos ou discretos, contudo, conforme modelo será necessário aplicar o método de discretização (I).

Uma vez com a base discretizada ocorre somente a separação dos clusters já classificados de acordo com a própria base de dados[3]. Isso é o funcionamento do fluxo (a), que nada mais é do que a separação da base em grupos que já classificados.

No passo (II) é onde serão executados os algoritmos de aprendizagem supervisionado, já visto nas sessões 1.1.1.1 e 1.1.1.2. Essa etapa é umas das mais importantes do método. O algoritmo supervisionado é aplicado várias vezes de acordo com o número de atributos do conjunto de dados, expresso no vetor de tamanho m. Onde o número de vezes será a quantidade de atributos da tabela, formando um conjunto de atributos importantes para cada grupo.

Seguindo para o processo (III) acontecerá a escolha do atributo mais relevante com seu valor mais frequente gerando rótulos para os grupos. Após essa etapa é criado um conjunto de rotulos para cada clusters. O fluxo (b) será utilizado caso houver algoritmo para ser executado.

2.3 Técnica de Correlação entre Atributos

Essa técnica [4] possui um grau de processamento diretamente proporcional a quantidade de características expressa na base de dados definido em R^m. Ela implica em

[3] UCI - Machine Learning Repository. http://archive.ics.uci.edu/ml/
[4] Extraída de (LOPES; MACHADO; RABELO,)

utilizar todos os atributos, menos o definido como classe, para fazer uma correlação entre eles junto ao algoritmo.

Pegando como exemplo uma base com os seguintes atributos: **atr1,atr2,atr3,classe**. Exclui o atributo classe, obtêm-se os 3($três$) primeiros atributos, onde cada um deles será utilizado como classe em refência aos outros atributos.

Em um primeiro processamento de três, o primeiro atributo **atr1** se torna classe e executado com os outros dois atributos restantes com um algoritmo supervisionado. O resultado da correlação entre os atributos **atr2, atr3** em relação ao **atr1**(figura 9) é armazanado em uma matriz, e logo depois é realizado com **atr2** sendo classe e assim sucessivamente até o último atributo.

Figura 9 – Exemplo da técnica aplicada ao atr1 sendo classe

2.4 Exemplo com Base Modelo

Para melhor esclarecer as etapas da figura 8, a tabela 2 contém uma base de dados que será utilizada para exemplificar todo o processo do modelo de resolução proposto nesta pesquisa. Logo na primeira coluna da tabela, retém um índice da linha da tabela responsável por identificar cada registro. Os outros campos são atributos que definem características do registro identificado pelo índice da primeira coluna.

Seguindo a definição 1 um elemento é expresso por um vetor de dimensão m, com tamanho igual ao número de atributos. Um exemplo do elemento 2 da tabela 2, pode ser representado por $\vec{e_2} = (1.26, 85.03, 20.45)$.

2.4.1 Processo (I) - Discretização

Segundo (CATLETT, 2006; HWANG; LI, 2002) o processo de discretização na etapa de treinamento pode aumentar a acurácia do algoritmo de aprendizado supervisionado. Dessa maneira a etapa de discretização ganha um papel importante no modelo, e também no processo de Rotulação (III), pois é utilizada uma inferência na faixa discretizada para encontrar o intervalo na faixa.

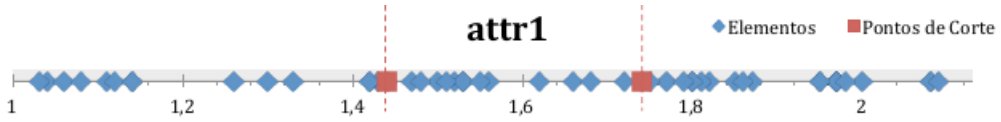
Para esse exemplo será utilizada a técnica de discretização por frequências iguais - EFD - e divisão de números de faixas, igual a R=3. Na figura[5] 10 poderá ser vizualizado

[5] Figura adaptada de (LOPES; MACHADO; RABELO,)

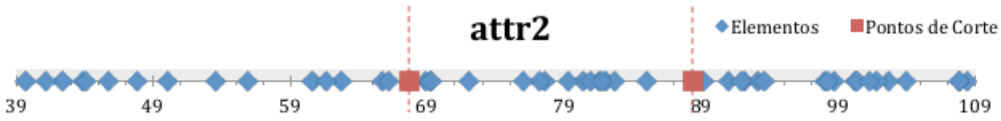
Tabela 2 – Base de Dados Modelo

	atr1	atr2	atr3	classe		atr1	atr2	atr3	classe
1	2.08	92.11	22.07	2	26	1.42	53.51	19.64	3
2	1.26	85.03	20.45	1	27	1.12	62.71	19.07	1
3	2.00	108.36	22.68	2	28	2.09	60.58	20.20	1
4	1.74	43.78	18.72	3	29	1.95	69.23	19.68	1
5	1.82	100.20	23.09	2	30	1.03	47.81	19.47	3
6	1.43	77.59	21.80	1	31	1.75	90.92	21.39	2
7	1.53	44.01	20.98	3	32	1.72	42.35	22.89	3
8	1.14	107.77	18.99	2	33	1.47	101.77	19.20	2
9	1.97	98.00	22.32	2	34	1.53	41.16	22.67	3
10	1.50	39.67	21.78	3	35	1.44	93.61	21.03	2
11	1.74	55.86	20.31	3	36	1.51	98.65	19.24	2
12	1.80	65.72	19.62	1	37	1.06	68.82	21.68	1
13	1.33	82.01	19.82	1	38	1.48	80.40	21.43	1
14	1.66	103.93	21.10	2	39	1.14	61.59	19.90	1
15	1.42	66.14	21.61	1	40	1.08	91.93	20.81	2
16	1.87	88.36	22.45	2	41	1.62	79.21	18.43	1
17	1.11	107.82	19.32	2	42	1.68	80.87	18.42	1
18	2.08	67.66	20.74	1	43	1.81	98.24	22.13	2
19	1.85	82.65	20.35	1	44	1.30	69.27	18.83	1
20	1.04	102.62	19.46	2	45	1.80	101.21	21.61	2
21	1.97	100.37	21.94	2	46	1.79	72.02	22.02	1
22	1.95	45.70	22.10	3	47	1.56	81.71	22.10	1
23	1.77	50.04	20.16	3	48	1.98	77.16	21.71	1
24	1.97	81.57	19.83	1	49	1.86	89.12	22.84	2
25	1.52	93.13	20.61	2	50	1.55	76.01	19.74	1

como é feita a discretização. Através da figura 10 fica claro o conteúdo da faixa 1, contendo



(a) Discretização em atr1



(b) Discretização em atr2

(c) Discretização em atr3

Figura 10 – Discretização de atributos utilizando EFD com R = 3

os valores iniciais até o primeiro ponto de corte, na faixa 2, são os valores após o primeiro

ponto de corte até o segundo ponto de corte. E na faixa 3 contém todos valores a partir do segundo ponto de corte.

Tabela 3 – Base de Dados Modelo Discretizada

	atr1	atr2	atr3	classe		atr1	atr2	atr3	classe
1	3	3	3	2	26	1	1	1	3
2	1	2	2	1	27	1	1	1	1
3	3	3	3	2	28	3	1	2	1
4	2	1	1	3	29	3	2	1	1
5	3	3	3	2	30	1	1	1	3
6	1	2	3	1	31	3	3	2	2
7	2	1	2	3	32	2	1	3	3
8	1	3	1	2	33	2	3	1	2
9	3	3	3	2	34	2	1	3	3
10	2	1	3	3	35	1	3	2	2
11	2	1	2	3	36	2	3	1	2
12	3	1	1	1	37	1	2	3	1
13	1	2	1	1	38	2	2	2	1
14	2	3	2	2	39	1	1	2	1
15	1	1	2	1	40	1	3	2	2
16	3	2	3	2	41	2	2	1	1
17	1	3	1	2	42	2	2	1	1
18	3	1	2	1	43	3	3	3	2
19	3	2	2	1	44	1	2	1	1
20	1	3	1	2	45	3	3	2	2
21	3	3	3	2	46	3	2	3	1
22	3	1	3	3	47	2	2	3	1
23	3	1	2	3	48	3	2	3	1
24	3	2	2	1	49	3	3	3	2
25	2	3	2	2	50	2	2	1	1

A tabela 3 é o resultado após a discretização de todos os atributos. Contudo sabe-se que ao se lidar com valores discretos onde cada intervalo representa uma faixa de valores poderá o algoritmo está perdendo um pouco de informação, mas por outro lado essa decisão tornará o aprendizado mais fácil de interpretar e com respostas mais rápidas.

2.4.2 Processo (II) - Algoritmos Supervisionados

Ao chegar nessa etapa, Processo (II) da figura 8, já se tem uma base discretizada e clusters formados, tabela 3. Agora é feita a execução do algorimo de aprendizado supervisionado e identificado os atributos de maior importância de cada cluster.

Uma vez com o conhecimento do cluster, serão percorridos todos os atributos, onde a cada iteração um atributo será a classe da vez. Nesse exemplo primeiramente o atributo **atr1** será classe, e os demais irão participar como entrada junto ao algoritmo, e verificar

seu grau de importância entre eles. Depois o atributo **atr2** irá ser classe, e depois o **atr3**, fechando o ciclo de todos os atributos do cluster. Como visualizado na figura 11

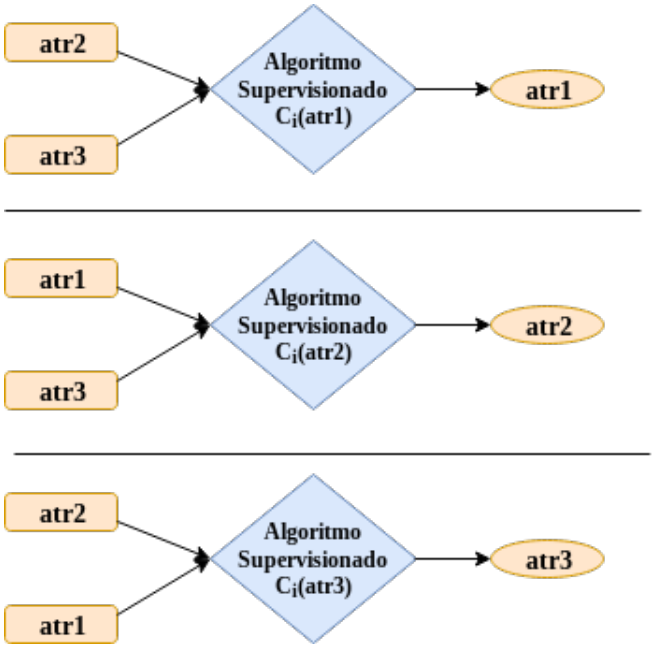
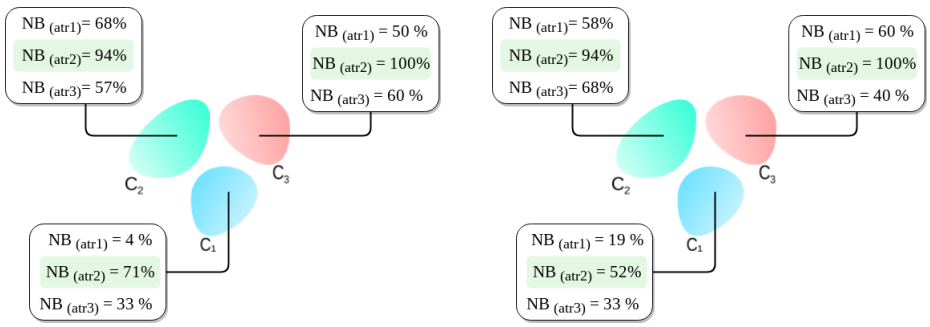


Figura 11 – Exemplo da técnica aplicada aos 3(*três*) atributos, cada um sendo classe em determinada iteração

Essa correlação entre os atributos junto com a aplicação dos algoritmos geram uma matriz de atributos importantes. O quão relevante o atributo será em relação ao cluster $c_i(atr)$, será dado em uma porcentagem de acerto quando aplicado como saída na execução de um algorimo supervisionado. Quanto maior sua porcentagem, mais correlacionado é o atributo em relação ao demais(figura 12), logo ele é considerado um atributo bem relevante. Sendo assim esse atributo poderá resumir as características do problema, podendo ser considerado um atributo importante e escolhido como rótulo.



(a) Execução do Naive Bayes

(b) Execução do algorimo de Árvore de Decisão

Figura 12 – Resultado dos Algoritmos

Na figura 12a mostra o resultado da execução do Naive Bayes em cima da Base Modelo e exibe os resultados em porcentagem de acerto de cada atributo em relação aos

demais. O mesmo acontece com a figura 12b onde é aplicado um algorimo de Árvore de Decisão, exibindo o resultado de todas as taxas de acerto, em porcentagem, dos atributos de seus respectivos clusters.

Uma forma de eliminar uma possível ambiguidade entre os clusters foi adicionar uma variável V. Com essa variável a seleção dos atributos rótulos de um clusters, seram todos os atributos que tiverem até uma diferença V em relação ao atributo de maior taxa de acerto, expresso em porcentagem. Portanto se o atributo de maior taxa de acerto possuir 90%, e o $V = 10\%$ então todos outros atributos que tiverem valores a partir de 80% seram selecionados como rótulo do cluster.

O valor da variável V é subjetivo e irá ser arbitrado de acordo com os resultados em cada aplicação do algoritmo em cima de um conjunto de dados. Nese exemplo os atributos importantes com $V = 12$ utilizando a figura 12a por cluster, teriam os rótulos $r_{c_i} : r_{c_1} = \{atr2\}$, $r_{c_2} = \{atr2\}$, $r_{c_3} = \{atr2\}$.

2.4.3 Processo (III) - Rotulação

Nesse processo de rotulação serão calculados os intervalos dos atributos que estão na figura 8 como Atributos Importantes, selecionados na etapa anterior. Para compor o rótulo r_{c_i} do cluster c_i é calculado a faixa do atributo que tiver maior frequência. É possível verificar neste exemplo, da Base Modelo, o resultado da figura 12a, onde o rótulo r_{c_1} é o **atr2=]67.66, 88.36]**, porque o valor da faixa de maior frequência do cluster c_1 em relação ao atributo **atr2** é a faixa 2(figura 10c), que é representa o limite inferior]67.66s e o limite superior, 88.36].

Uma vez terminado o processo (III) de rotulação, o fluxo *b* do modelo da figura 8, só é seguido caso seja necessário para executar outro algoritmo.

Contudo pode-se definir os rótulos nesta etapa da seguinte maneira:

- Algoritmo Naive Bayes 12a aplicado na BD Modelo
  $r_{c_1} = (atr2,]67.66, 88.36])$;
  $r_{c_2} = (atr2,]88.36, 108.36])$;
  $r_{c_3} = (atr2, [39.67, 67.66])$;
- Algoritmo de Árvore de Decisão 12b aplicado na BD Modelo
  $r_{c_1} = (atr2,]67.66, 88.36])$;
  $r_{c_2} = (atr2,]88.36, 108.36])$;
  $r_{c_3} = (atr2, [39.67, 67.66])$;

Logo abaixo o algoritmo 2 exibe a rotina em forma de pseudocódigo para melhor entendimento.

Algorithm 2: Rotina de Rotulação

```
1  Carrega_valores_auxiliares(V, R, TipoDiscretização);
2  Carrega_BD;
3  Discretiza_BD;
4  Separa_em_clusters_de_acordo_com_classificação_BD;
5  while existir clusters do
6      while existir atributos do
7          prepara_vetor_atributos/classe;
8          Aplica_algoritmo_supervisionado;
9          Calcula_matriz_de_porcentagem_de_acertos;
10     Carrega_atributos_importantes_considerando_V;
11     Associa_valores_aos_intervalos;
12 Exibe_rótulos_todos_clusters;
```

3 Resultados

Os resultados obtidos aqui neste capítulo foram referentes a aplicação do método de rotulação em 3(*três*) bases de dados distintas. Um dos primeiros passos na análise de aprendizagem de máquina é quando o analista prepara os dados para poder utilizar um método de aprendizagem apropriado.

Então a escolha da base de dados também tem influência direta em bons resultados. E sabendo disso a escolha dos conjuntos de dados utilizados nesta pesquisa foi por conta delas apresentarem características diferentes, e também por serem conhecidas, facilitando a análise e servindo de amostra a outras base.

3.1 Implementação

Para conseguir gerar os resultados aqui escritos foram feitas implementações utilizando a ferramenta MATLAB [1], onde junto a ela é possível utilizar suas funções de aprendizado de máquina já prontas. MATLAB possui uma linguagem técnica, e de fácil implementação por já possuir uma gama de funções[2] preparadas para aprendizado de máquina. Por esses motivos essa ferramenta foi escolhida para colocar em prática essa pesquisa.

Foram realizados vários testes com o intuito de tentar otimizar resultados e poder comparálos a outras pesquisas já escritas. Seguindo essa linha foi determinado a escolha de 3(*três*) bases de dados já conhecidas, onde na implementação de cada uma delas surgiu algumas alterações, dependendo da base, na variável(V), quantidade de faixas(R) e método de discretização(EWD,EFD). Essas mudanças para cada base servirão para otimizar os resultados.

Cada base de dados será aplicado dois algoritmos de aprendizado supervisionado que possuem paradigmas diferentes para servir de amostra e poder assim tirar conclusões sobre a rotulação em quaisquer algoritmos supervisionados.

Os algoritmos utilizados foram o Naive Bayes, sessão 1.1.1.2, com paradigma estatístico. E também o algoritmo Classification e Regression Trees - CART, 1.1.1.1, com paradigma simbólico de árvore de decisão.

[1] http://www.mathworks.com/products/matlab/
[2] versão: R2016a(9.0.0.341360); 64-bit (glnxa64)

3.2 Seeds - Identificação de Tipos de Semente

Essa base foi extraída da UCI Machine Learning[3], composta por 7(*sete*) atributos definindo suas características e mais uma definindo sua classificação, sendo este último um atributo classe responsável por identificar o tipo de semente. Possuindo um total de 210 registros classificados em 3(*três*) categorias:

- 70 elementos do tipo Kama;
- 70 elementos do tipo Rosa;
- 70 elementos do tipo Canadian.

Na configuração de implementação foi utilizado o método EFD de discretização com divisão em três faixas, $R = 3$ para todos os atributos, e inserido o valor de variação $V = 3\%$.

Na tabela 4 e tabela 7 são apresentados os resultados com a execução do algoritmo Naive Bayes e CART respectivamente. Elas são formadas por uma coluna informando os **Clusters**, **Rótulos** compostos pelo **Atributo** e sua **Faixa** de valor. Junto também a coluna **Relevância** exibindo a resposta do algoritmo em porcentagem, da correlação do atributo em relação aos outros atributos do cluster, retirado da tabela 5 e da tabela 8 respectivamente. E por último a coluna **Elem Fora da Faixa** que mostra a quantidade de elementos que não estão dentro da faixa do rótulo. Essa última coluna tem a função de exibir em números a quantidade de valores que não estão participando da porcentagem da coluna de **Relevância**. Para quem está analisando a tabela é interessante mais essa dado, pois pode compara com o total de elementos do grupo.

3.2.1 Naive Bayes

Tabela 4 – Resultado da aplicação do algoritmo Naive Bayes

	Rótulos			
Cluster	Atributos	Faixa	Relevância(%)	Elem fora da Faixa
1	area	] 12.78 ∼ 16.14]	92%	14
2	area	] 16.14 ∼ 21.18]	95%	6
	lkernel	] 5.826 ∼ 6.675]	92%	6
3	perimetro	[12.41 ∼ 13.73]	95%	5

Analisando a coluna rótulo da tabela 4, nota-se que o atributo **area** aparece tanto no cluster 1 como também no cluster 2. A técnica envolve não só o rótulo como também a faixa que os valores mais se repetem dentro do atributo. Nesse caso pode-se observar que o atributo se repete entre os clusters. Mas no cluster 1, a faixa de valores difere do cluster 2, sem comentar que no cluster 2 existe outro atributo compondo o rótulo, **lkernel**.

[3] http://archive.ics.uci.edu/ml/

A seleção dos atributos rótulos acontece da diferença da variável $V = 3\%$ em relação ao atributo de maior relevência. Caso essa variável tenha o valor alterado, os rótulos dos clusters poderão sofrer mudanças, pois poderá aumentar ou diminuir o número de atributos dos rótulos, dependendo do valor inserido em V. Através da tabela 5 é possível analisar todos os valores de relevância gerados para os atributos e analisar qual valor pode-se inserir em V para montar o rótulo.

Tabela 5 – Resultado da Correlação dos atributos pelo Naive Bayes; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	92.8	87.1	50.0	75.7	85.7	60.0	65.7
	2	95.7	91.4	47.1	92.8	90.0	28.5	85.7
	3	91.4	95.7	71.4	85.7	91.4	64.2	58.5

A tabela 5 é formada por clusters representado pelas linhas, e colunas representado por atributos. Essa tabela é fruto da implementação do Naive Bayes em cima dessa base de dados, e foi gerada para auxiliar a retirada dos atributos rótulos. Uma análise pode ser feita através desses dados e ajudar a definir um valor para a variável V. Percebe-se que algumas características são mais bem correlacionadas que outras, através de seus valores mais altos. Isso indica o grau de relacionamento entre os atributos após a aplicação do algoritmo.

Tabela 6 – Resultado de 4(*quatro*) execuções do algoritmo Naive Bayes; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	92.8	87.1	48.5	77.1	82.8	57.1	65.7
	2	94.2	90.0	45.7	92.8	90.0	38.5	87.1
	3	91.4	95.7	72.8	85.7	91.4	64.2	60.0

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	92.8	87.1	47.1	77.1	87.1	60.0	65.7
	2	94.2	90.0	47.1	92.8	91.4	32.8	87.1
	3	91.4	95.7	72.8	85.7	92.8	64.2	60.0

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	94.2	85.7	48.5	77.1	82.8	61.4	65.7
	2	92.8	90.0	50.0	92.8	90.0	32.8	87.1
	3	91.4	95.7	72.8	85.7	92.8	64.2	60.0

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	91.4	88.5	54.2	75.7	85.7	62.8	61.4
	2	95.7	90.0	50.0	92.8	90.0	38.5	85.7
	3	91.4	95.7	72.8	85.7	94.2	64.2	57.1

Para conseguir ter uma idéia mais ampla dessas informações, na tabela 6 é exposto o resultado de 4(*quatro*) execuções do Algoritmo Naive Bayes, e pode-se constatar que mesmo havendo algumas alterações em seus valores nos atributos em cada execução, a correlação entre os atributos não oferece muita alteração. Como exemplo, o atributo **area**, possui o melhor grau de correlacionamento em seu grupo, mesmo testado em quatro execuções, como mostrado na tabela 6.

Segue abaixo o resultado do algoritmo Naive Bayes na base de dados **Seeds** com seus rótulos:

- $r_{c_1} = \{(area,]12.78\ 16.14])\}$
- $r_{c_2} = \{(area,]16.14\ 21.18]), (Lkernel,]5.826\ 6.675])\}$
- $r_{c_3} = \{(perimetro, [12.41\ 13.73])\}$

3.2.2 CART

Já na tabela 7, tem-se o resultado da aplicação de outro algoritmo supervisionado, mas dessa vez com paradigma simbólico e não mais no estatístico. No MATLAB o algoritmo de árvore de decisão utilizado é o CART.

Tabela 7 – Resultado da aplicação do algoritmo CART

	Rótulos			
Cluster	Atributos	Faixa	Relevância(%)	Elem fora da Faixa
1	area	] 12.78 ∼ 16.14]	91%	14
	perimetro	[13.73 ∼ 15.18]	94%	14
2	area	] 16.14 ∼ 21.18]	95%	6
	perimetro	] 15.18 ∼ 17.25]	98%	7
3	perimetro	[12.41 ∼ 13.73]	95%	5
	wkernel	[2.63 ∼ 3.049]	97%	9

Pode-se verificar na tabela 7 que os clusters 1 e 2 possuem o mesmo conjunto de atributos selecionados no campo de rótulo. Mas isso não implica dizer que os dois grupos são identificados pelo mesmo rótulo. O rótulo é composto pelos atributos e pelas faixas, onde a faixa é escolhida é a faixa onde se tem o maior número de valores que se repetem nessa faixa. Então, caso exista um vetor de elementos já discretizados, $\vec{e}_{(c_i)} = \{1, 1, 1, 2, 2, 2, 2, 3, 3\}$. Neste vetor o valor que mais se repete é o 2 , então a faixa 2 foi a que mais se repetiu e com isso é a escolhida para compor o rótulo com o atributo mais relevante.

Para entender a escolha desses atributos no campo de rótulos, a tabela 8 exibe o resultado gerado na execução do algoritmo em cima da base. Cada valor desses é o resultado da aplicação do algoritmo enquanto o atributo era a classe da vez coforme figura 11, sessão 2.4.2. O atributo de maior valor junto com os atributos da diferença de V com o mais relevante, são escolhidos para ser rótulos. Na linha(cluster) 1 o maior valor é o atributo perimetro. Pega o valor encontrado em perimetro, e subtrai de $V = 3$. A partir daí o(s) atributo(s) que possuí(rem) um valor que está entre este resultado até o mais alto, irá compor o rótulo.

Foram realizadas vários teste, onde alguns deles estão na tabela 9. Essas operações foram execuções do algoritmo CART em cima da base, para provar que a técnica de correlação de atributos, 2.3 é funcional para este algoritmo. O mesmo pode ser visto

Tabela 8 – Resultado da Correlação dos atributos pelo CART; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	91.4	94.2	58.5	80.0	81.4	61.4	61.4
	2	98.5	98.5	51.4	90.0	88.5	42.8	88.5
	3	92.7	95.7	80.0	88.5	97.1	58.5	78.5

no algoritmo de paradigma estatístico 3.2.1 realizado nessa pesquisa. O comportamento de ambos foram bem semelhantes, pois eles seguem o padrão de valores os quais não se alteram muito a cada iteração.

Tabela 9 – Resultado de 4(*quatro*) iterações do algoritmo CART; Legenda dos Atributos: (A)area, (B)perimetro, (C)compacteness, (D)Lkernel, (E)Wkernel, (F)asymetry, (G)lkgroove

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	91.4	94.2	58.5	80.0	74.2	55.7	60.0
	2	98.5	98.5	50.0	90.0	88.5	41.4	90.0
	3	92.8	95.7	80.0	88.5	97.1	55.7	77.1

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	91.4	94.2	62.8	78.5	81.4	61.4	57.1
	2	98.5	98.5	54.2	90.0	88.5	40.0	90.0
	3	92.8	95.7	80.0	88.5	97.1	60.0	77.14

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	93.8	93.6	61.8	83.2	89.2	53.2	71.0
	2	98.2	98.3	61.9	93.0	90.5	25.2	90.1
	3	95.5	96.3	82.4	90.9	97.7	59.3	77.0

		Atributos						
		A	B	C	D	E	F	G
Clusters	1	92.8	94.2	60.0	80.0	84.2	64.2	60.0
	2	98.5	98.5	47.1	91.4	90.0	42.8	88.5
	3	91.4	95.7	80.0	88.5	97.1	55.7	77.1

O resultado do algoritmo CART na base de dados **Seeds** tem como rótulos:

- $r_{c_1} = \{(area,]12.78\ 16.14]), (perimetro,]13.73\ 15.18])\}$
- $r_{c_2} = \{(area,]16.14\ 21.18]), (perimetro,]15.18\ 17.25])\}$
- $r_{c_3} = \{(perimetro, [12.41\ 13.73]), (wkernet, [2.63\ 3.049])\}$

3.3 Iris - Identificação de Tipos de Plantas

3.4 Glass - Identificação de Tipos de Vidro

Conclusões e Trabalhos Futuros

Proin non sem. Donec nec erat. Proin libero. Aliquam viverra arcu. Donec vitae purus. Donec felis mi, semper id, scelerisque porta, sollicitudin sed, turpis. Nulla in urna. Integer varius wisi non elit. Etiam nec sem. Mauris consequat, risus nec congue condimentum, ligula ligula suscipit urna, vitae porta odio erat quis sapien. Proin luctus leo id erat. Etiam massa metus, accumsan pellentesque, sagittis sit amet, venenatis nec, mauris. Praesent urna eros, ornare nec, vulputate eget, cursus sed, justo. Phasellus nec lorem. Nullam ligula ligula, mollis sit amet, faucibus vel, eleifend ac, dui. Aliquam erat volutpat.

Conclusões

Fusce vehicula, tortor et gravida porttitor, metus nibh congue lorem, ut tempus purus mauris a pede. Integer tincidunt orci sit amet turpis. Aenean a metus. Aliquam vestibulum lobortis felis. Donec gravida. Sed sed urna. Mauris et orci. Integer ultrices feugiat ligula. Sed dignissim nibh a massa. Donec orci dui, tempor sed, tincidunt nonummy, viverra sit amet, turpis. Quisque lobortis. Proin venenatis tortor nec wisi. Vestibulum placerat. In hac habitasse platea dictumst. Aliquam porta mi quis risus. Donec sagittis luctus diam. Nam ipsum elit, imperdiet vitae, faucibus nec, fringilla eget, leo. Etiam quis dolor in sapien porttitor imperdiet.

Cras pretium. Nulla malesuada ipsum ut libero. Suspendisse gravida hendrerit tellus. Maecenas quis lacus. Morbi fringilla. Vestibulum odio turpis, tempor vitae, scelerisque a, dictum non, massa. Praesent erat felis, porta sit amet, condimentum sit amet, placerat et, turpis. Praesent placerat lacus a enim. Vestibulum non eros. Ut congue. Donec tristique varius tortor. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nam dictum dictum urna.

Phasellus vestibulum orci vel mauris. Fusce quam leo, adipiscing ac, pulvinar eget, molestie sit amet, erat. Sed diam. Suspendisse eros leo, tempus eget, dapibus sit amet, tempus eu, arcu. Vestibulum wisi metus, dapibus vel, luctus sit amet, condimentum quis, leo. Suspendisse molestie. Duis in ante. Ut sodales sem sit amet mauris. Suspendisse ornare pretium orci. Fusce tristique enim eget mi. Vestibulum eros elit, gravida ac, pharetra sed, lobortis in, massa. Proin at dolor. Duis accumsan accumsan pede. Nullam blandit elit in magna lacinia hendrerit. Ut nonummy luctus eros. Fusce eget tortor.

Trabalhos Futuros

Ut sit amet magna. Cras a ligula eu urna dignissim viverra. Nullam tempor leo porta ipsum. Praesent purus. Nullam consequat. Mauris dictum sagittis dui. Vestibulum sollicitudin consectetuer wisi. In sit amet diam. Nullam malesuada pharetra risus. Proin lacus arcu, eleifend sed, vehicula at, congue sit amet, sem. Sed sagittis pede a nisl. Sed tincidunt odio a pede. Sed dui. Nam eu enim. Aliquam sagittis lacus eget libero. Pellentesque diam sem, sagittis molestie, tristique et, fermentum ornare, nibh. Nulla et tellus non felis imperdiet mattis. Aliquam erat volutpat.

Referências

BARBER, D. *Bayesian Reasoning and Machine Learning*. [s.n.], 2011. ISSN 9780521518147. ISBN 9780511804779. Disponível em: <http://ebooks.cambridge.org/ref/id/CBO9780511804779>. Citado 3 vezes nas páginas 1, 4 e 8.

CATLETT, J. Into Ordered Discrete Attributes. v. 3, n. 1989, p. 2006, 2006. Citado 2 vezes nas páginas 8 e 16.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and Unsupervised Discretization of Continuous Features. *Machine Learning Proceedings 1995*, v. 0, p. 194–202, 1995. ISSN 0717-6163. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/B9781558603776500323>. Citado na página 8.

FILHO, V. P. R. *Rotulacao de grupos utilizando conjuntos fuzzy*. Tese (Doutorado) — Universidade Federal do Piauí, 2015. Citado 3 vezes nas páginas 15, 11 e 12.

HWANG, G. J.; LI, F. A Dynamic Method for Discretization of Continuous Attributes. *Lecture Notes in Computer Science - Intelligent Data Engineering and Automated Learning - IDEAL 2002: Third International Conference*, v. 2412/2002, p. 506, 2002. ISSN 16113349. Disponível em: <http://www.springerlink.com/content/4n05b2n6x0cx4tlk>. Citado 2 vezes nas páginas 8 e 16.

KOTSIANTIS, S.; KANELLOPOULOS, D. Discretization Techniques : A recent survey. *GESTS International Transactions on Computer Science and Engineering*, v. 32, n. 1, p. 47–58, 2006. Citado na página 8.

LIMA, B. V. A. Método Semissupervisionado de Rotulação e Classificação Utilizando Agrupamento por Sementes e Classificadores. 2015. Citado na página 11.

LOPES, L. A.; MACHADO, V. P.; RABELO, R. D. A. L. Automatic Labeling of Groupings through Supervised Machine Learning. Citado 6 vezes nas páginas 15, 10, 11, 13, 14 e 16.

LUCCA, G. et al. Uma implementação do algoritmo Naïve Bayes para classificação de texto. *Centro de Ciências Computacionais - Universidade Federal do Rio Grande (FURG) Rio Grande - RS - Brasil*, p. 1–4, 2013. Citado na página 7.

MADUREIRA, D. F. *Analise de sentimento para textos curtos*. Tese (Doutorado) — Fundacao Getulio Vargas, Rio de Janeiro, 2017. Citado na página 7.

MCCALLUM, A.; NIGAM, K. A Comparison of Event Models for Naive Bayes Text Classification. 1997. Citado na página 7.

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. 432 p. ISSN 10450823. ISBN 9781577354260. Citado 2 vezes nas páginas 1 e 3.

RAIMUNDO, L. R.; MATTOS, M. C. D.; WALESKA, P. O Algoritmo de Classificação CART em uma Ferramenta de Data Mining. 2008. Citado na página 6.

RUSSEL, S.; NORVIG, P. *Inteligência Artificial*. 3ª. ed. Rio de Janeiro: [s.n.], 2013. ISBN 9780136042594. Citado 2 vezes nas páginas 3 e 4.

WU, X. et al. *Top 10 algorithms in data mining*. [S.l.: s.n.], 2008. v. 14. 1–37 p. ISSN 02191377. ISBN 1011500701. Citado na página 7.

YOHANNES, Y.; WEBB, P. *Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity*. International Food Policy Research Institute, 1999. (Microcomputers in policy research). ISBN 9780896293373. Disponível em: <https://books.google.com.br/books?id=7iuq4ikyNdoC>. Citado na página 6.

Apêndices

APÊNDICE A – Primeiro Apêncice

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.

APÊNDICE B – Perceba que o texto do título desse segundo apêndice é bem grande

Maecenas dui. Aliquam volutpat auctor lorem. Cras placerat est vitae lectus. Curabitur massa lectus, rutrum euismod, dignissim ut, dapibus a, odio. Ut eros erat, vulputate ut, interdum non, porta eu, erat. Cras fermentum, felis in porta congue, velit leo facilisis odio, vitae consectetuer lorem quam vitae orci. Sed ultrices, pede eu placerat auctor, ante ligula rutrum tellus, vel posuere nibh lacus nec nibh. Maecenas laoreet dolor at enim. Donec molestie dolor nec metus. Vestibulum libero. Sed quis erat. Sed tristique. Duis pede leo, fermentum quis, consectetuer eget, vulputate sit amet, erat.

Donec vitae velit. Suspendisse porta fermentum mauris. Ut vel nunc non mauris pharetra varius. Duis consequat libero quis urna. Maecenas at ante. Vivamus varius, wisi sed egestas tristique, odio wisi luctus nulla, lobortis dictum dolor ligula in lacus. Vivamus aliquam, urna sed interdum porttitor, metus orci interdum odio, sit amet euismod lectus felis et leo. Praesent ac wisi. Nam suscipit vestibulum sem. Praesent eu ipsum vitae pede cursus venenatis. Duis sed odio. Vestibulum eleifend. Nulla ut massa. Proin rutrum mattis sapien. Curabitur dictum gravida ante.

Phasellus placerat vulputate quam. Maecenas at tellus. Pellentesque neque diam, dignissim ac, venenatis vitae, consequat ut, lacus. Nam nibh. Vestibulum fringilla arcu mollis arcu. Sed et turpis. Donec sem tellus, volutpat et, varius eu, commodo sed, lectus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Quisque enim arcu, suscipit nec, tempus at, imperdiet vel, metus. Morbi volutpat purus at erat. Donec dignissim, sem id semper tempus, nibh massa eleifend turpis, sed pellentesque wisi purus sed libero. Nullam lobortis tortor vel risus. Pellentesque consequat nulla eu tellus. Donec velit. Aliquam fermentum, wisi ac rhoncus iaculis, tellus nunc malesuada orci, quis volutpat dui magna id mi. Nunc vel ante. Duis vitae lacus. Cras nec ipsum.

Anexos

ANEXO A – Nome do Primeiro Anexo

Sed mattis, erat sit amet gravida malesuada, elit augue egestas diam, tempus scelerisque nunc nisl vitae libero. Sed consequat feugiat massa. Nunc porta, eros in eleifend varius, erat leo rutrum dui, non convallis lectus orci ut nibh. Sed lorem massa, nonummy quis, egestas id, condimentum at, nisl. Maecenas at nibh. Aliquam et augue at nunc pellentesque ullamcorper. Duis nisl nibh, laoreet suscipit, convallis ut, rutrum id, enim. Phasellus odio. Nulla nulla elit, molestie non, scelerisque at, vestibulum eu, nulla. Ut odio nisl, facilisis id, mollis et, scelerisque nec, enim. Aenean sem leo, pellentesque sit amet, scelerisque sit amet, vehicula pellentesque, sapien.

ANEXO B – Nome de Outro Anexo

Phasellus id magna. Duis malesuada interdum arcu. Integer metus. Morbi pulvinar pellentesque mi. Suspendisse sed est eu magna molestie egestas. Quisque mi lorem, pulvinar eget, egestas quis, luctus at, ante. Proin auctor vehicula purus. Fusce ac nisl aliquam ante hendrerit pellentesque. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi wisi. Etiam arcu mauris, facilisis sed, eleifend non, nonummy ut, pede. Cras ut lacus tempor metus mollis placerat. Vivamus eu tortor vel metus interdum malesuada.