Comparação de Técnicas de Aprendizado utilizando o WEKA

Equipe:

Amilton Fontoura de Camargo Junior

Kauane Larisse de Oliveira Benitis

Sumário

Introdução

O que é Inteligência Artificial (IA)?

- Inteligência artificial é um termo abrangente que engloba o Aprendizado de Máquina
- Também sob esse termo genérico está o Aprendizado Profundo (AP), um subconjunto do aprendizado de máquina que utiliza modelos de redes neurais para executar tarefas como reconhecimento de imagem e processamento de linguagem.

Introdução

A inteligência artificial propicia diversas melhorias em áreas distintas.

- Ao medir a biometria no esporte
- Ajudam os agricultores a saber quando regar as plantações para obter colheitas excelentes
- Permitem que os meteorologistas calculem o degelo.
- Cidades inteligentes utilizam dados para o gerenciamento de energia,
- Detectar doenças, realizar sequenciamento genômico e acompanhar tratamentos.

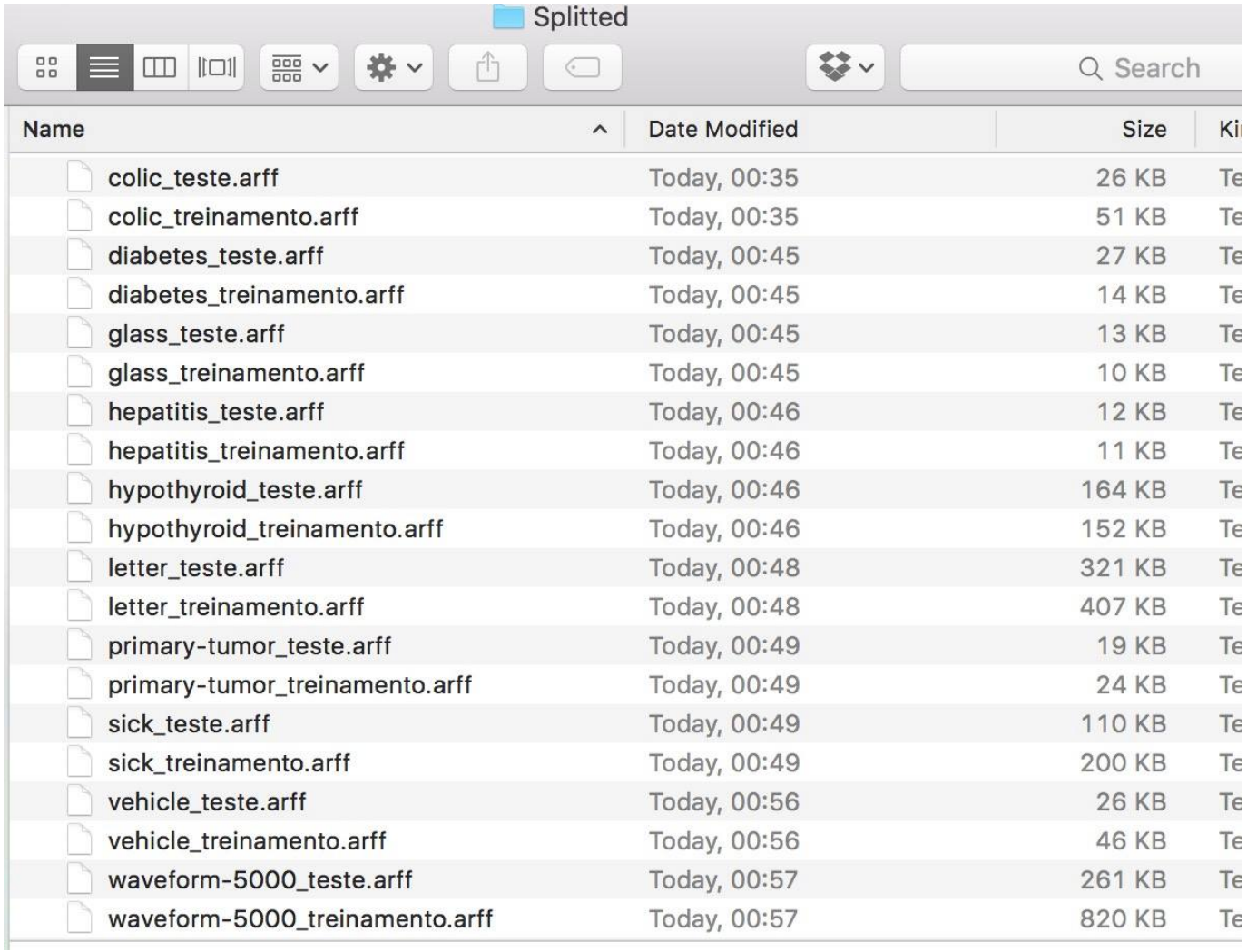
Conjunto de Dados

Nome do Conjunto	Numero de Amostras	Numero de Classes	Numero de Atributos
Colic	63983	368	23
Dermatology	32417	366	35
Diabetes	37419	768	9
Glass	17823	214	10
Hepatitis	17135	155	20
Hypothyroid	310897	3772	30
Iris	7486	150	4
Primary Tumor	34090	339	18
Vehicle	63838	4	18
Tae	4120	151	6

Split


Split - Dados
Split de Dados
Arquivo de entrada:
Procurar
Arquivos de saída
Treinamento:
Padrão
Teste:
Padrão
Divisão:
Treinamento: 20 % e Teste: 80 %
Escolher as amostras aleatoriamente
Processar
Limpar campos
Tae

Split


Splitted
Search
Name
Date Modified
Size
colic_teste.arff Today, 00:35 26 KB
colic_treinamento.arff Today, 00:35 51 KB
diabetes_teste.arff Today, 00:45 27 KB
diabetes_treinamento.arff Today, 00:45 14 KB
glass_teste.arff Today, 00:45 13 KB
glass_treinamento.arff Today, 00:45 10 KB
hepatitis_teste.arff Today, 00:46 12 KB
hepatitis_treinamento.arff Today, 00:46 11 KB
hypothyroid_teste.arff Today, 00:46 164 KB
hypothyroid_treinamento.arff Today, 00:46 152 KB
letter_teste.arff Today, 00:48 321 KB
letter_treinamento.arff Today, 00:48 407 KB
primary-tumor_teste.arff Today, 00:49 19 KB
primary-tumor_treinamento.arff Today, 00:49 24 KB
sick_teste.arff Today, 00:49 110 KB
sick_treinamento.arff Today, 00:49 200 KB
vehicle_teste.arff Today, 00:56 26 KB
vehicle_treinamento.arff Today, 00:56 46 KB
waveform-5000_teste.arff Today, 00:57 261 KB
waveform-5000_treinamento.arff Today, 00:57 820 KB

Técnicas Utilizadas

Amostragem

- Cross Validation (10 folds)

Técnicas de Aprendizado Supervisionadas

- J48
- NaiveBayes
- Multilayer Perceptron
- IBk

Técnica de Aprendizado Não-Supervisionada

- kMeans

Análise e Discussão dos Resultados — Obtidos

Conjunto	Custo (s)							
	Cross Validation	NaiveBayes	J48	MLP	IBK 1	IBK 2	Kmeans 1	Kmeans 2
Colic	0.00	0.01	0.01	5.51	0.00	0.00	0.11	0.07
Dermatology	0.00	0.00	0.00	23.56	0.00	0.00	0.11	0.11
Diabetes	0.00	0.00	0.02	0.57	0.00	0.00	0.27	0.15
Glass	0.00	0.00	0.01	0.38	0.00	0.00	0.01	0.02
Hepatitis	0.00	0.00	0.00	0.32	0.00	0.00	0.04	0.02
Hypothyroid	0.00	0.01	0.02	21.92	0.00	0.00	7.78	7.51
Iris	0.00	0.00	0.00	0.09	0.00	0.00	0.01	0.01
Primary Tumor	0.00	0.00	3.74	3.76	0.00	0.00	0.11	0.06
Vehicle	0.00	0.00	0.10	2.12	0.00	0.00	0.6	0.62
Tae	0.00	0.00	0.00	0.10	0.00	0.00	0.01	0.01
Média	0.00	0.00	0.39	5.83	0.00	0.00	0.91	0.86
Desvio Padrão	0.00	0.00	1.18	9.10	0.00	0.00	2.42	2.34

Análise e Discussão dos Resultados — Obtidos

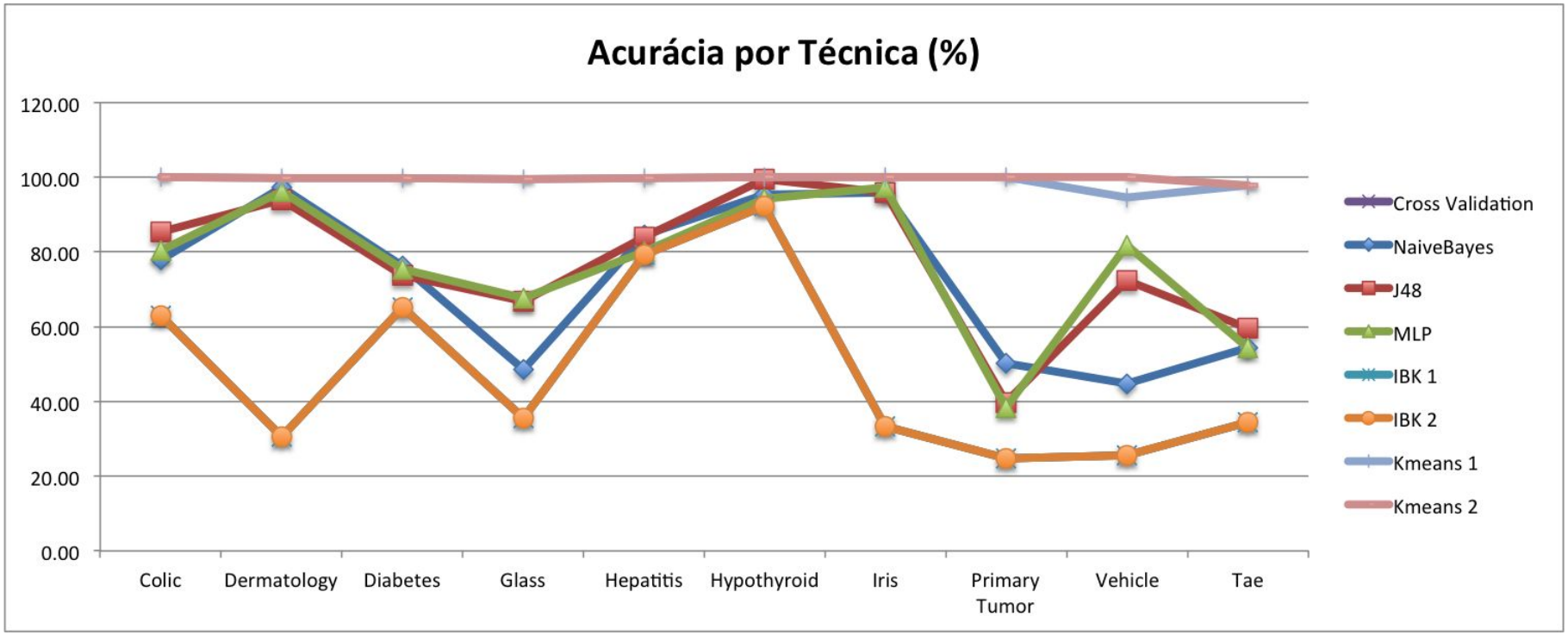
Conjunto	Acurácia (%)							
	Cross Validation	NaiveBayes	J48	MLP	IBK 1	IBK 2	Kmeans 1	Kmeans 2
Colic	63.04	77.99	85.33	80.43	63.04	63.04	100.00	100.00
Dermatology	30.60	97.27	93.99	96.17	30.60	30.60	99.80	99.80
Diabetes	65.10	76.30	73.83	75.39	65.10	65.10	99.80	99.80
Glass	35.51	48.60	66.82	67.76	35.51	35.51	99.50	99.50
Hepatitis	79.35	84.52	83.87	80.00	79.35	79.35	99.70	99.70
Hypothyroid	92.29	95.28	99.58	94.17	92.29	92.29	100.00	100.00
Iris	33.33	96.00	96.00	97.33	33.33	33.33	100.00	100.00
Primary Tumor	24.78	50.15	39.82	38.35	24.78	24.78	100.00	100.00
Vehicle	25.65	44.80	72.46	81.68	25.65	25.65	94.60	100.00
Tae	34.44	54.30	59.60	54.30	34.44	34.44	97.80	97.80
Média	48.41	72.52	77.13	76.56	48.41	48.41	99.12	99.66
Desvio Padrão	24.40	21.19	18.54	18.90	24.40	24.40	1.72	0.68

Análise e Discussão dos Resultados — Obtidos

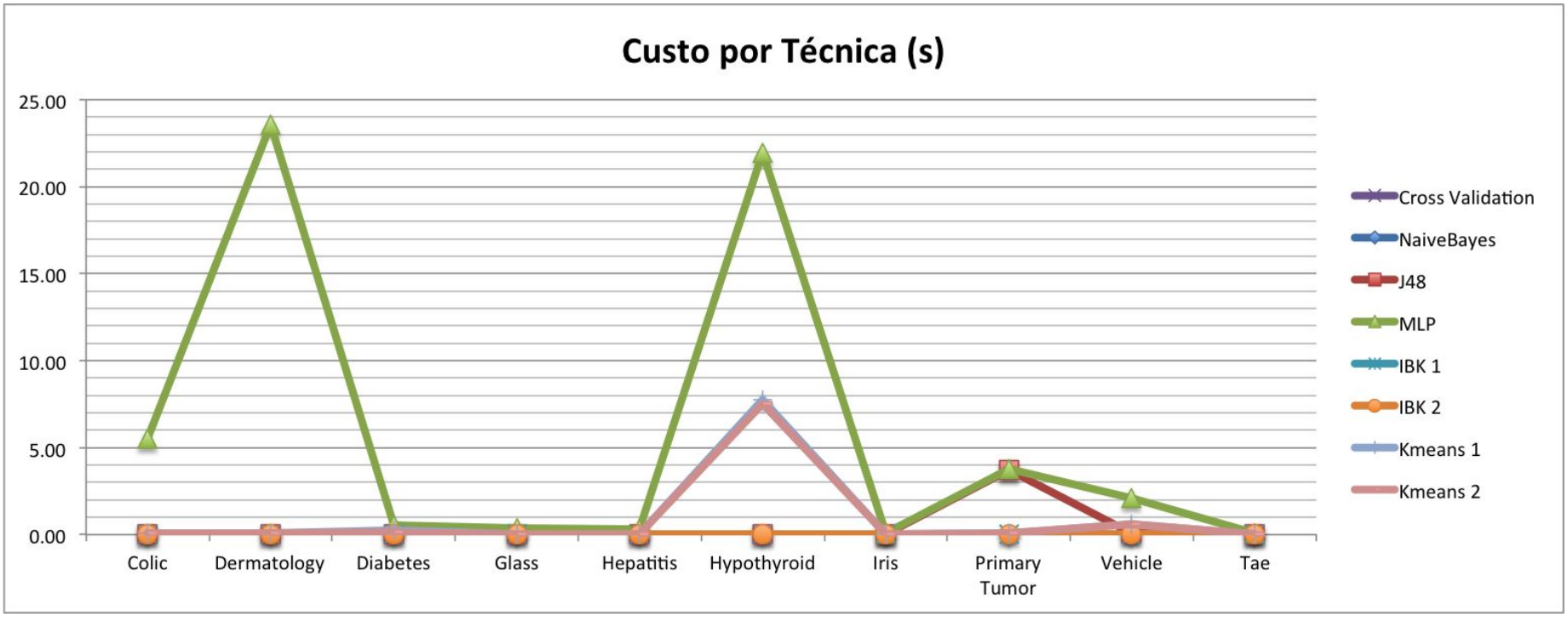
Conjunto	Média Acurácia	Desvio Padrão Acurácia	Média Custo	Desvio Padrão Custo
Colic	79.11	15.53	0.71	1.94
Dermatology	72.35	34.63	2.97	8.32
Diabetes	77.55	14.52	0.13	0.20
Glass	61.09	27.11	0.05	0.13
Hepatitis	85.73	8.86	0.05	0.11
Hypothyroid	95.74	3.58	4.66	7.78
Iris	73.67	33.43	0.01	0.03
Primary Tumor	50.33	31.95	0.96	1.72
Vehicle	58.81	32.03	0.43	0.73
Tae	58.39	26.36	0.02	0.03

Análise e Discussão dos Resultados — Obtidos

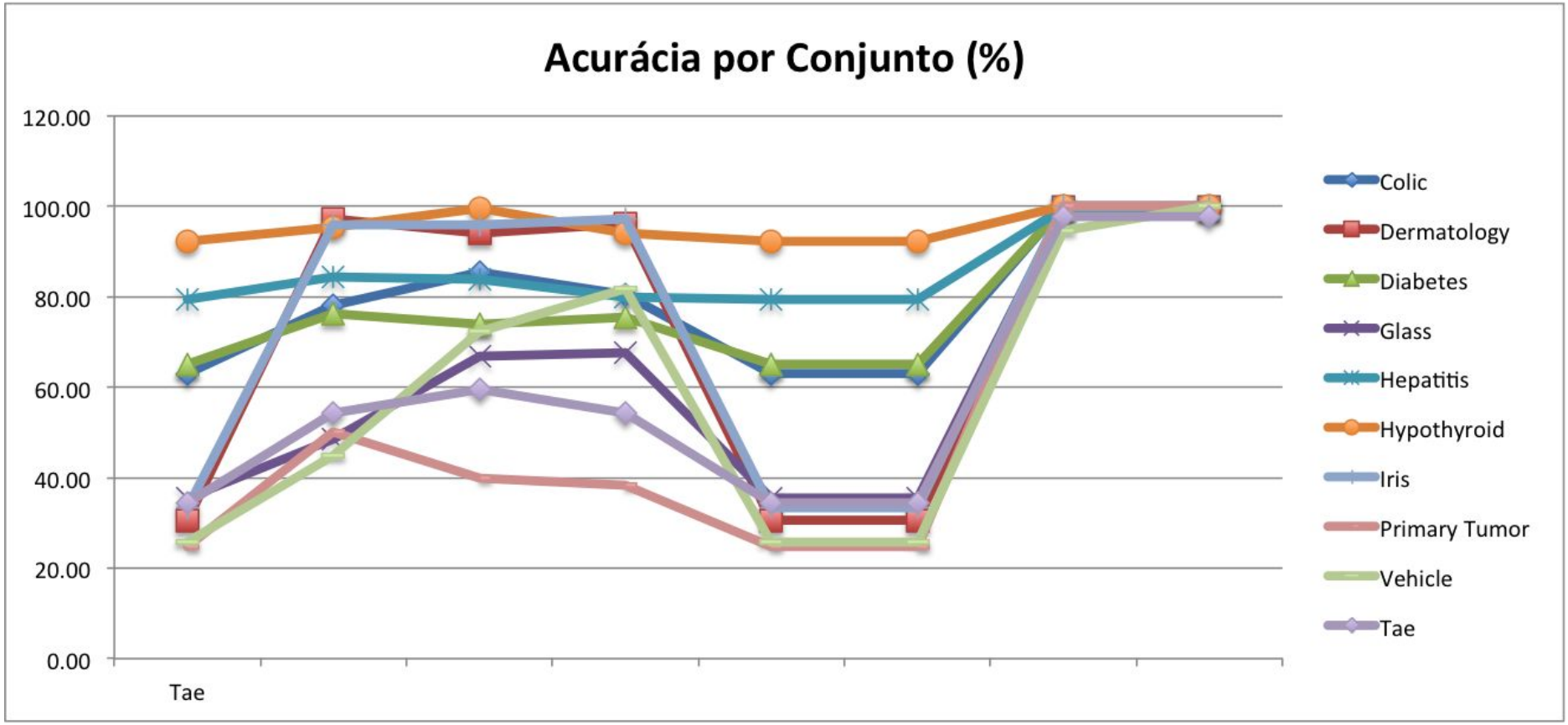
Análise e Discussão dos Resultados — Obtidos



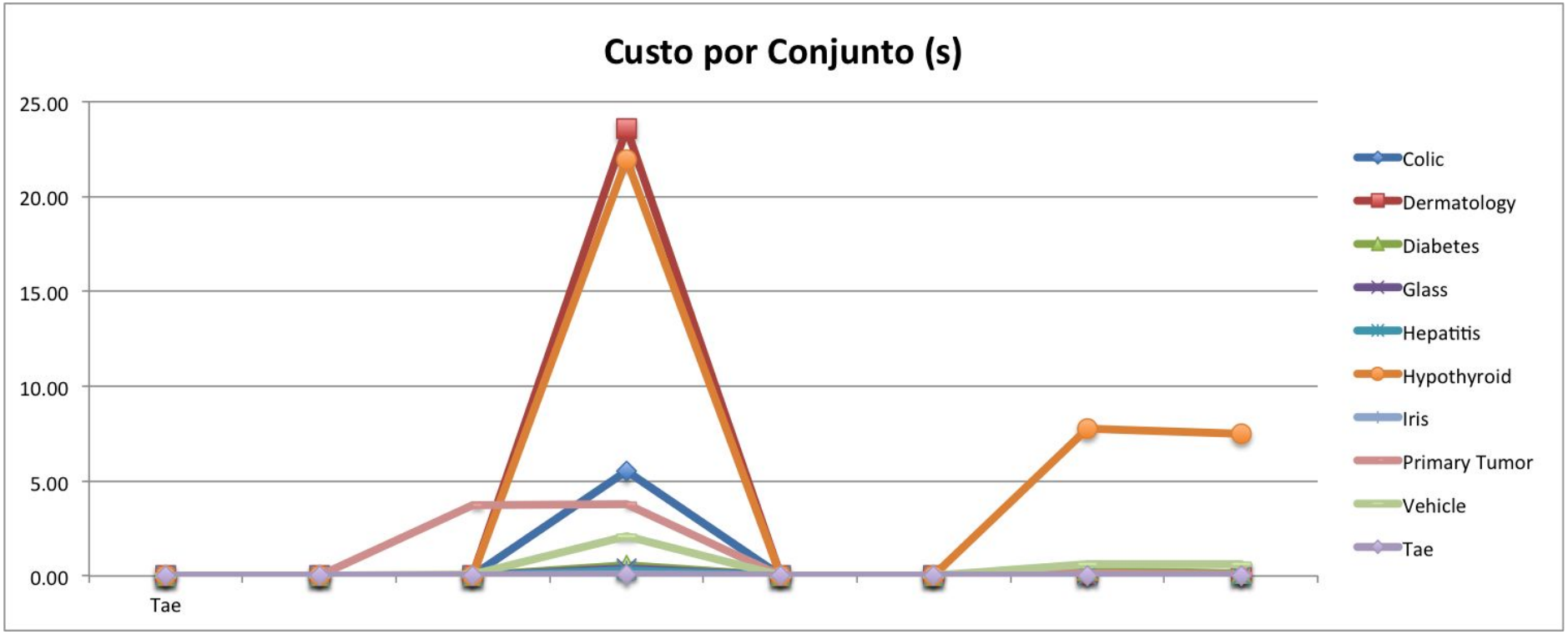
Análise e Discussão dos Resultados Obtidos



Análise e Discussão dos Resultados — Obtidos



Análise e Discussão dos Resultados — Obtidos



Conjunto - Split - Escolha

Hypothyroid

- Grande número de amostras - 310897
- Um dois mais custosos antes do Split com uma média de 4,66 segundos
- Maior índice de Acurácia dos conjuntos com uma média de 95,74%

Análise e Discussão dos Resultados — Obtidos - Split

Hypothyroid - J48 - Acurácia: 99,80% / 99,47%

Técnica: J48 - Training				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 3013**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.84%	0.07%	0.00%	0.00%
compensated_hypothyroid	0.00%	5.54%	0.00%	0.00%
primary_hypothyroid	0.03%	0.07%	2.42%	0.00%
secondary_hypothyroid	0.03%	0.00%	0.00%	0.00%

Técnica: J48 - Test				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 755**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	93.77%	0.00%	0.00%	0.00%
compensated_hypothyroid	0.00%	3.58%	0.00%	0.00%
primary_hypothyroid	0.26%	0.13%	2.12%	0.00%
secondary_hypothyroid	0.13%	0.00%	0.00%	0.00%

Análise e Discussão dos Resultados — Obtidos - Split

Hypothyroid - NaiveBayes - Acurácia: 99,80% / 99,47%

Técnica: NaiveBayes - Training				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 3013**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.11%	0.50%	0.27%	0.03%
compensated_hypothyroid	3.62%	1.86%	0.07%	0.00%
primary_hypothyroid	0.27%	0.20%	2.06%	0.00%
secondary_hypothyroid	0.00%	0.00%	0.00%	0.03%

Técnica: NaiveBayes - Test				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 755**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	93.11%	0.13%	0.07%	0.07%
compensated_hypothyroid	0.46%	1.72%	0.00%	0.00%
primary_hypothyroid	0.00%	0.26%	2.25%	0.00%
secondary_hypothyroid	0.13%	0.00%	0.00%	0.00%

Análise e Discussão dos Resultados — Obtidos - Split

Hypothyroid - MLP - Acurácia: 99,80% / 99,47%

Técnica: MLP - Training				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 3013**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.54%	0.27%	0.10%	0.00%
compensated_hypothyroid	3.09%	1.99%	0.46%	0.00%
primary_hypothyroid	0.07%	0.07%	2.39%	0.00%
secondary_hypothyroid	0.00%	0.00%	0.03%	0.00%

Técnica: MLP - Test				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 755**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	92.19%	1.32%	0.26%	0.00%
compensated_hypothyroid	2.25%	0.93%	0.40%	0.00%
primary_hypothyroid	0.00%	0.13%	2.38%	0.00%
secondary_hypothyroid	0.00%	0.00%	0.13%	0.00%

Análise e Discussão dos Resultados — Obtidos - Split

Hypothyroid - IBk=2 - Acurácia: 99,80% / 99,47%

Técnica: IBk k = 2 - Training				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 3013**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.90%	0.00%	0.00%	0.00%
compensated_hypothyroid	4.38%	1.16%	0.00%	0.00%
primary_hypothyroid	0.80%	0.27%	1.46%	0.00%
secondary_hypothyroid	0.00%	0.03%	0.00%	0.00%

Técnica: IBk k = 2 - Test				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 755**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	93.38%	0.40%	0.00%	0.00%
compensated_hypothyroid	3.31%	0.26%	0.00%	0.00%
primary_hypothyroid	1.46%	0.53%	0.53%	0.00%
secondary_hypothyroid	0.13%	0.00%	0.00%	0.00%

Análise e Discussão dos Resultados — Obtidos - Split

Hypothyroid - IBk=5 - Acurácia: 99,80% / 99,47%

Técnica: IBk k = 5 - Training				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 3013**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	91.77%	0.13%	0.00%	0.00%
compensated_hypothyroid	4.61%	0.90%	0.03%	0.00%
primary_hypothyroid	1.16%	0.17%	1.19%	0.00%
secondary_hypothyroid	0.03%	0.00%	0.00%	0.00%

Técnica: IBk k = 5 - Test				
Conjunto: hypothyroid.arff	**Classe: Class**	**Amostras: 755**		
↓ Real Escolhido →	negative	compensated_hypothyroid	primary_hypothyroid	secondary_hypothyroid
negative	93.25%	0.53%	0.00%	0.00%
compensated_hypothyroid	3.05%	0.53%	0.00%	0.00%
primary_hypothyroid	1.59%	0.00%	0.93%	0.00%
secondary_hypothyroid	0.13%	0.00%	0.00%	0.00%

Trabalho Correlato

- *Lexicographic preferences for predictive modeling of human decision making: A new machine learning method with an application in accounting*

- Ou traduzindo: **Preferências lexicográficas para a modelagem preditiva da tomada de decisão humana: Um novo método de aprendizagem de máquina com aplicação em contabilidade**

- De acordo com os autores, uma estratégia de tomada de decisão humana é a heurística "Escolher a Melhor", sem integrar todas as informações, apenas de forma "competitiva" escolher a aparentemente melhor.
- Outra estratégia é o modelo lexicográfico, que é comparar duas opções, verificando um critério entre eles que difere e usando o mesmo como fator de decisão, sem considerar os demais.

Trabalho Correlato

- **O que há de diferente?**
- Ao invés de comparar cada atributo separadamente, faz a comparação de todos eles em simultâneo, aproximando-se da heurística humana.

- Outra diferença proposta pelos autores é um método para discretizar os domínios de atributos numéricos, para análise par a par, facilitando a aplicação da heurística proposta.

- Em um estudo de caso, o artigo investiga um problema importante e altamente complexo de decisão do mundo real com relação à elaboração de relatórios financeiros sobre modelo de pensões profissionais (contribuinte/não-contribuinte, segurado/não-segurado, etc.)

Trabalho Correlato

- Da mesma forma que outras decisões de negócios, a administração da empresa tem que considerar muitos fatores (atributos) ao decidir sobre um método, e pode ser uma boa alternativa aplicar uma heurística lexicográfica.

- Possui **Lista de Preferências Lexicográficas (LPs)** com as alternativas a serem consideradas, onde o algoritmo faz o aprendizado destas listas para, então, tomar as decisões.

$$\boldsymbol{x} \in \mathcal{X} = \mathcal{D}(V) = \mathcal{D}(A_1) \times \cdots \times \mathcal{D}(A_n),$$

Onde V são as alternativas e D os domínios dos atributos.

Trabalho Correlato

$\mathcal{T} = \left\{(\boldsymbol{x}_i^*, \boldsymbol{x}_i)\right\}_{i=1}^{N}$

Conjunto de treinamento

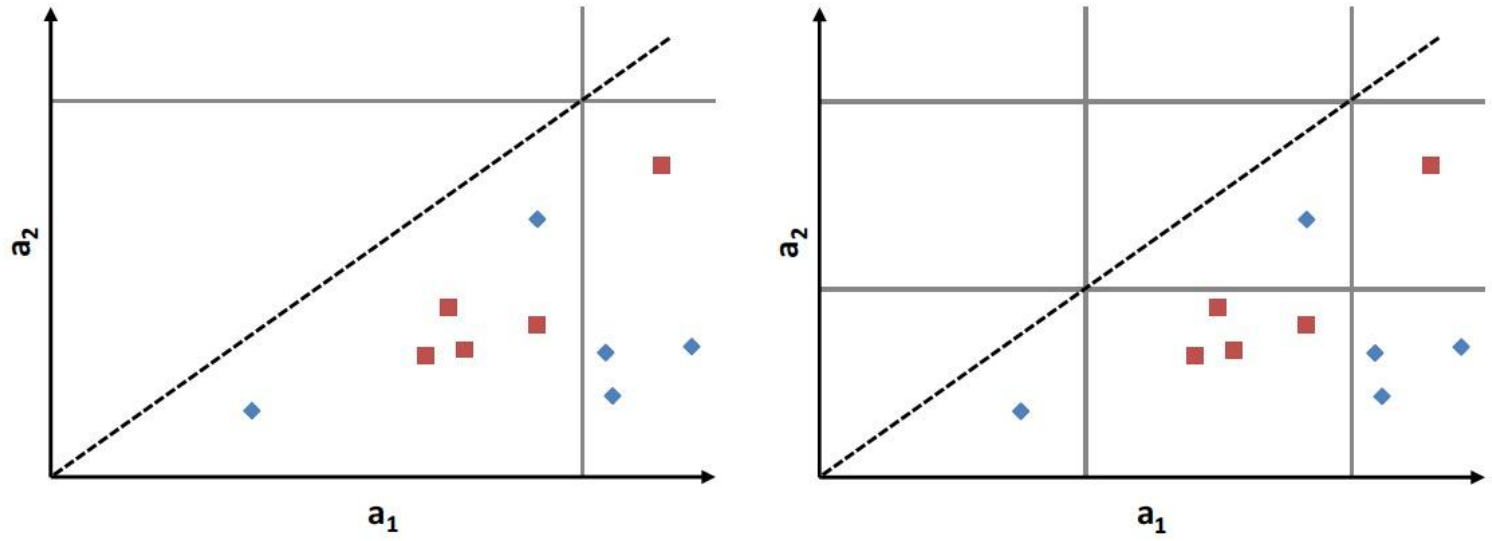
Algoritmo para expansão dos pares de escolha mais prováveis

Algorithm 1: LPLL.

Input : training data $\mathcal{T}$, set of attributes V, maximal grouping size g_{max}, maximum partition size r_{max}
Output: LP list l

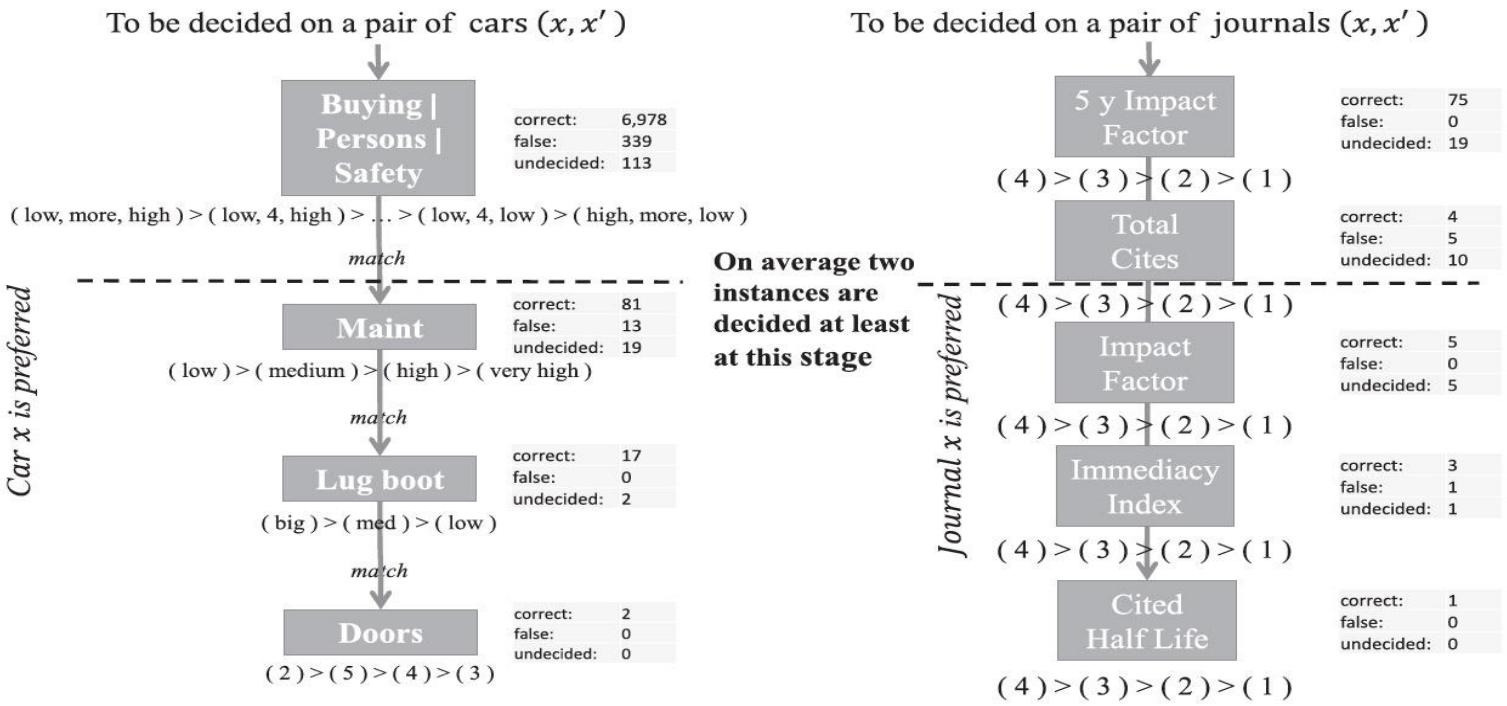
$l \leftarrow \emptyset$, $V' \leftarrow V$, $\mathcal{I}' \leftarrow \{1, \ldots, n\}$
while $\mathcal{T} \neq \emptyset$ *and* $V' \neq \emptyset$ **do**
  $I' \leftarrow \emptyset$, $CR \leftarrow 0$, $CP \leftarrow 0$
  for $\mathcal{I} \subseteq \mathcal{I}'$, $|\mathcal{I}| \leq g_{max}$ **do**
    discretize $A_i, i \in \mathcal{I}$ into at most r_{max} bins
    determine $\sqsupset_{\mathcal{I}}$ on $\mathcal{D}(V_{\mathcal{I}})$ maximally consistent with $\mathcal{T}$
    compute $\mathrm{CR}(\sqsupset_{\mathcal{I}}, \mathcal{T})$ and $\mathrm{CP}(\sqsupset_{\mathcal{I}}, \mathcal{T})$
    if $\mathrm{CR}(\sqsupset_{\mathcal{I}}, \mathcal{T}) = \mathrm{CR}$ && $\mathrm{CP} < \mathrm{CP}(\sqsupset_{\mathcal{I}}, \mathcal{T})$ **then**
      $\mathrm{CP} \leftarrow \mathrm{CP}(\sqsupset_{\mathcal{I}}, \mathcal{T})$
      $I' \leftarrow \mathcal{I}$
    if $\mathrm{CR}(\sqsupset_{\mathcal{I}}, \mathcal{T}) > \mathrm{CR}$ **then**
      $\mathrm{CR} \leftarrow \mathrm{CR}(\sqsupset_{\mathcal{I}}, \mathcal{T})$
      $\mathrm{CP} \leftarrow \mathrm{CP}(\sqsupset_{\mathcal{I}}, \mathcal{T})$
      $I' \leftarrow \mathcal{I}$
    reverse discretization for $A_i, i \in \mathcal{I}$
  discretize $A_i, i \in I'$ into at most r_{max} bins
  $\mathcal{I}' \leftarrow \mathcal{I}' \backslash I'$
  $V' \leftarrow V' \backslash V_{I'}$
  remove every $(\vec{x}, \vec{x}') \in \mathcal{T}$ decided by $\sqsupset_{I'}$
  add item $(V_{I'}, \sqsupset_{I'})$ to l

Trabalho Correlato



Agrupamento/expansão dos atributos de decisão. À esquerda, apenas uma divisão e na direita duas divisões, onde cada divisão possui elementos "puros", sendo eles do tipo vermelho (negativos) ou do tipo azul (positivos).

Trabalho Correlato



Exemplo: decisão entre um par de carros e um par de artigos, considerando sempre dois atributos

Considerações Finais

- O aprendizado supervisionado possui maior acurácia, por terem especificadas as classes.
- Aprendizado não-supervisionado é melhor quando não se há muita informação a respeito do conjunto de dados a ser analisado.
- Mais atributos/dimensões resultam numa melhor classificação, porém deve-se atentar à maneira com que seleciona-se tais dados.
- Técnicas de aprendizado ativo não foram citadas, mas podem fazer muita diferença quando os conjuntos de dados são muito grandes.
- Se o conjunto de dados é muito grande, o computador que está executando o experimento deve conter muita memória RAM! As técnicas de aprendizado facilitam na quantidade das classificações, mas têm um grande custo computacional.
- A acurácia aumenta ao dividir os conjuntos de dado em teste e treinamento.

Referências

BARANAUSKAS, J. A. **Aprendizado de máquinas: Conceitos e definições.** 2007. Disponível em <http://dcm.ffclrp.usp.br/augusto/teaching/ami/AM-I-Conceitos-Definicoes.pdf>. Acesso em 10 de Novembro de 2016.

BRÄUNING, Michael, HÜLLERMEIER, Eyk, KELLER, Tobias, GLAUM, Martin. **Lexicographic preferences for predictive modeling of human decision making: A new machine learning method with an application in accounting.** ELSEVIER: European Journal of Operational Research. Publicado em 18 de Agosto de 2016. Disponível em <http://www.sciencedirect.com/science/article/pii/S0377221716306944>. Acesso em 10 de Novembro de 2016.

OLIVEIRA, Cristiano. **Inteligência Artificial – O que é?.** Laboratório de Estruturas e Materiais Estruturais - USP. Disponível em <http://www.lem.ep.usp.br/Pef411/~Cristiano%20Oliveira/CristianoOliveira/Paginas/InteligenciaArtificial.htm>. Acesso em 10 de Novembro de 2016.