

Aprendizado Multi-objetivo

Usando conflitos para aprender

Marcos M. Raimundo

3 de Setembro de 2021 - Campinas - Brazil

Introdução

Problemas de classificação

Exemplos de problemas de classificação:

- Detecção de spam.
- Identificação de espécie/gênero.
- Identificação de impressão digital.
- Detecção de crises de epilepsia.
- Acesso a crédito.
- Reincidência em crime.

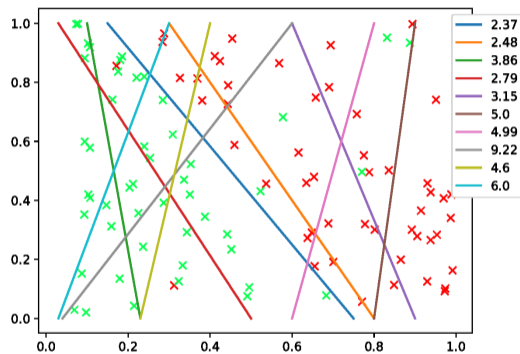


Figura 1: Representação de classificadores e suas perdas logísticas.

N amostras:

- Dados de entrada: $\mathbf{x}_i \in \mathbb{R}^d : i \in \{1, \dots, N\}$;
- Dados de saída classificação binária: $y_i \in \{0, 1\} : i \in \{1, \dots, N\}$.
- Dados de saída classificação multiclasse:
 $y_i^k \in \{0, 1\} : i \in \{1, \dots, N\}, k \in \{1, \dots, K\}, \sum_k y_i^k = 1$.

Usando de exemplo a regressão logística multinomial, vamos considerar um modelo

$f_k(\theta, \mathbf{x}) = \frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}}$ para estimar a pertinência de $\mathbf{x}_i \in \mathbb{R}^d$ à classe k , e $\theta \in \mathbb{R}^d$ é o vetor de pesos que queremos ajustar para encontrar o melhor modelo.

Com isso, chegamos a $\sum_k l_k(\theta) + \lambda r(\theta) \equiv \sum_{k=1}^K - \left[\sum_{i=1}^N y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] + \lambda r(\theta)$.

Complexidade do modelo – Redes Neurais

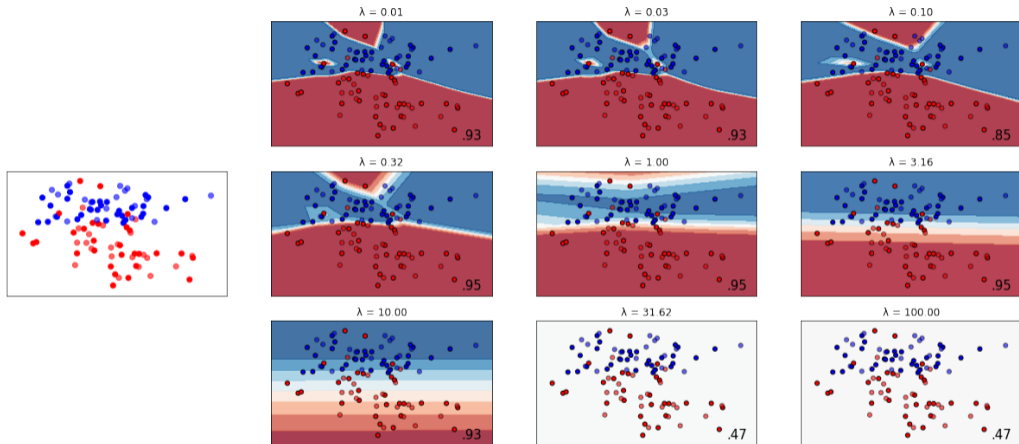


Figura 2: Representação dos dados de classificação e das fronteiras de decisão.

Complexidade vs. erro de aprendizado do modelo

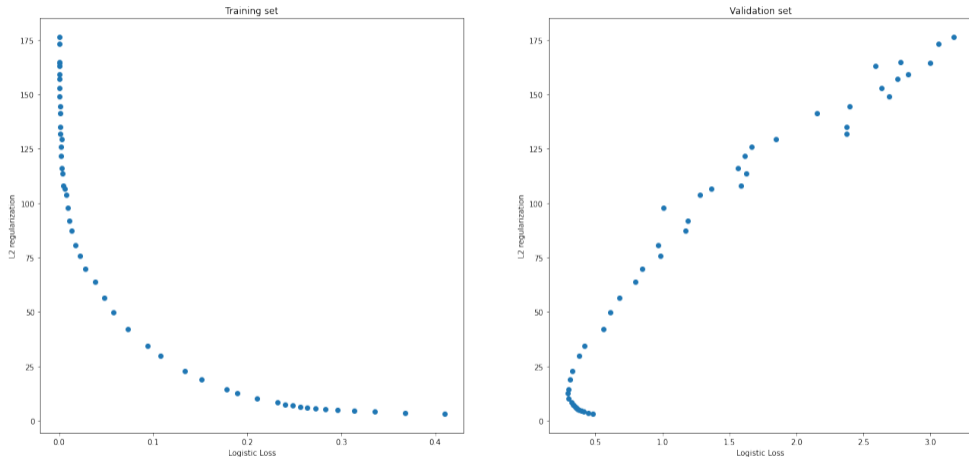
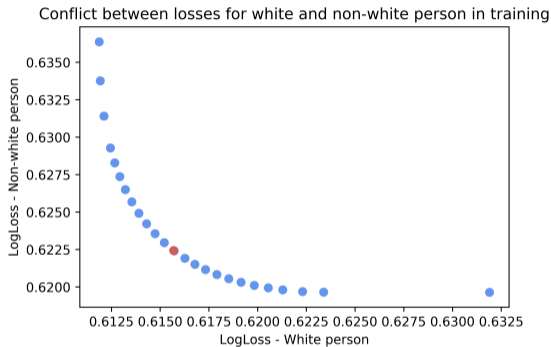
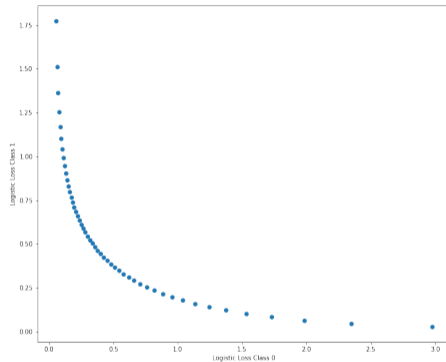


Figura 3: Relação entre perda de aprendizado e complexidade do modelo nos dados de treinamento e validação.

Outros conflitos no aprendizado



(a) Aprendizado com discriminação



(b) Classes desbalanceadas

Figura 4: Fronteira de Pareto para problemas de aprendizado de máquina.

Conflito no aprendizado – Classificação Multiclasse

$$\min_{\theta} \sum_{k=1}^K - \left[\sum_{i=1}^N y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] \equiv \sum_{k=1}^K l_k(\theta). \quad (1)$$

Problemas:

- Preferência por classes com mais amostras.
- Expectativas do usuário.

Caso médico:

- Não diagnosticar um paciente em crise (FN).
- Diagnosticar estado normal como crise (FP).

Otimização multiobjetivo

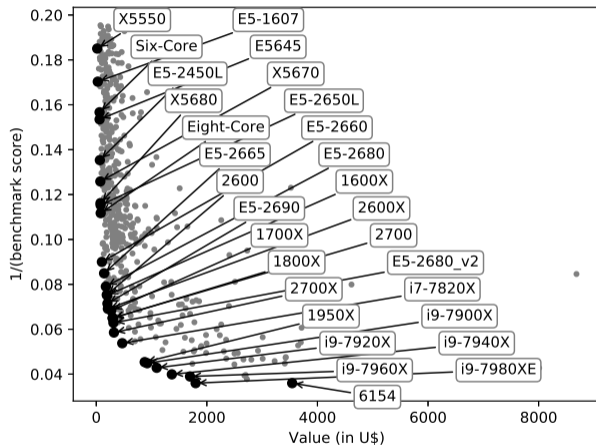


Figura 5: Representação multiobjetivo de CPUs de alto desempenho cpubenchmark.net com 485 CPUs e somente 29 CPUs Pareto ótimas.

- Problemas com metas contraditórias (exemplo de CPU).
- Resolvemos escolhendo o melhor custo-benefício (preferências implícitas).
- Solução dominada (dominância).
- Solução eficiente ou Pareto-ótima.
- Fronteira de Pareto.

Explorando conflitos em aprendizado de máquina

1. **Modelagem multiobjetivo:** objetivos conflitantes são formalizados explicitamente e agregados numa formulação multiobjetivo;
2. **Treinamento multiobjetivo:** usando essas modelagens, são usados métodos de otimização multiobjetivo a posteriori (NISE [Cohon et al., 1979] e MONISE [Raimundo and Von Zuben, 2017]) para encontrar as soluções Pareto-ótimas mais representativas.
3. **Tomada de decisão a posteriori:** dado que é construído um conjunto de soluções, nós podemos selecionar a melhor solução, ou podemos usar esses candidatos Pareto-ótimos como componentes de um ensemble, explorando a literatura de ensemble para agregar esses modelos visando encontrar a melhor performance de aprendizado.

Otimização Multiobjetivo

Definição

O problema multiobjetivo é definido por:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{f}(\mathbf{x}) \equiv \{f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})\} \\ \text{sujeito a} \quad & \mathbf{x} \in \Omega, \Omega \in \mathbb{R}^n \\ & \mathbf{f}(\mathbf{x}) : \Omega \rightarrow \Psi \in \mathbb{R}^m \\ & f_i(\mathbf{x}) : \Omega \rightarrow \mathbb{R}, i = 1, 2, \dots, m. \end{aligned} \tag{2}$$

Representação de um problema multiobjetivo

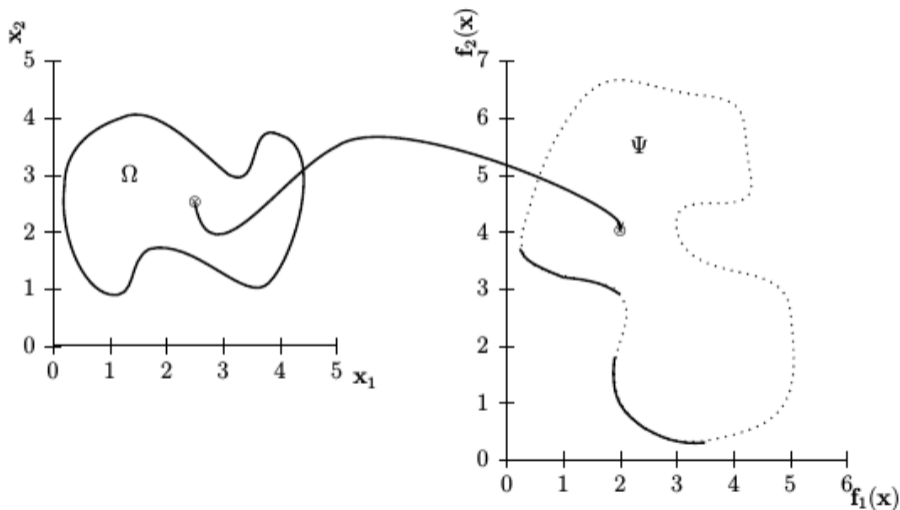


Figura 6: Representação do espaço de decisão e do espaço dos objetivos.

Definição

A definição do método da soma ponderada (também chamado de escalarização por pesos) é dada por:

$$\begin{aligned} \min_x \quad & \mathbf{w}^\top \mathbf{f}(\mathbf{x}) \\ \text{sujeito a} \quad & \mathbf{x} \in \Omega, \\ & \mathbf{f}(\mathbf{x}) : \Omega \rightarrow \Psi, \Omega \subset \mathbb{R}^n, \Psi \subset \mathbb{R}^m \\ & \mathbf{w} \in \mathbb{R}^m, w_i \geq 0 \forall i \in \{1, 2, \dots, m\}. \end{aligned} \tag{3}$$

Método da soma ponderada

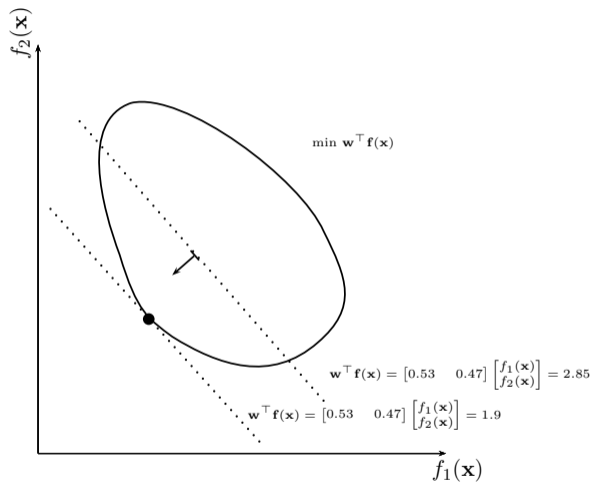


Figura 7: Representação de uma solução produzida pelo método da soma ponderada [Raimundo et al., 2020].

$$\begin{aligned}
 \min_{\mathbf{w}, \underline{\mathbf{r}}, \bar{\mathbf{r}}} \quad & -\mu = \mathbf{w}^\top \underline{\mathbf{r}} - \mathbf{w}^\top \bar{\mathbf{r}} \\
 \text{s.a.} \quad & \mathbf{w}^i \underline{\mathbf{r}} \geq \mathbf{w}^i \mathbf{f}(x^i) \quad \forall i \in P \\
 & \mathbf{w}^\top \bar{\mathbf{r}} \leq \mathbf{w}^\top \mathbf{f}(x^i) \quad \forall i \in P \\
 & \underline{\mathbf{r}} \geq \mathbf{z}^{\text{utopian}}, \underline{\mathbf{r}} \leq \bar{\mathbf{r}}, \\
 & \mathbf{w} \geq \mathbf{0}, \mathbf{w}^\top \mathbf{1} = 1.
 \end{aligned} \tag{4}$$

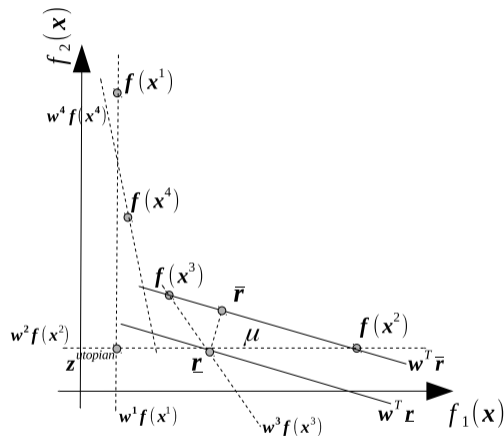


Figura 8: Representação do procedimento de otimização do algoritmo MONISE para encontrar o próximo vetor de pesos \mathbf{w} .

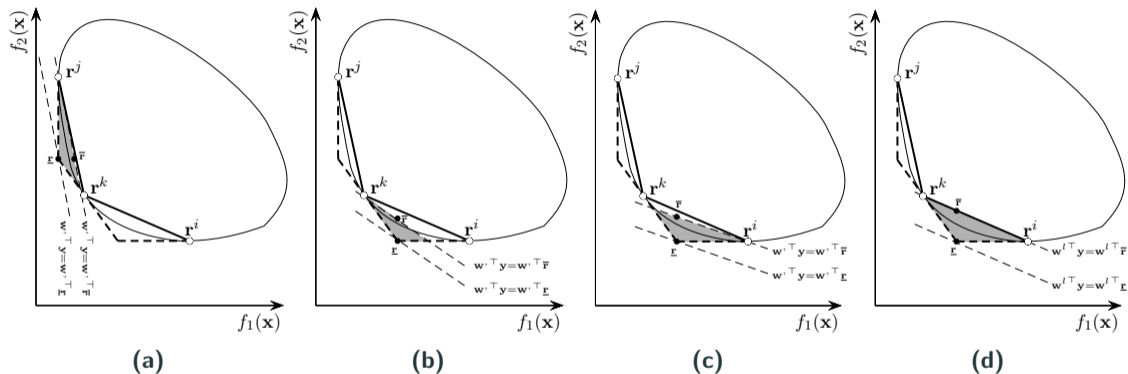


Figura 9: Soluções sub-ótimas para o cálculo do vetor de pesos do algoritmo MONISE.

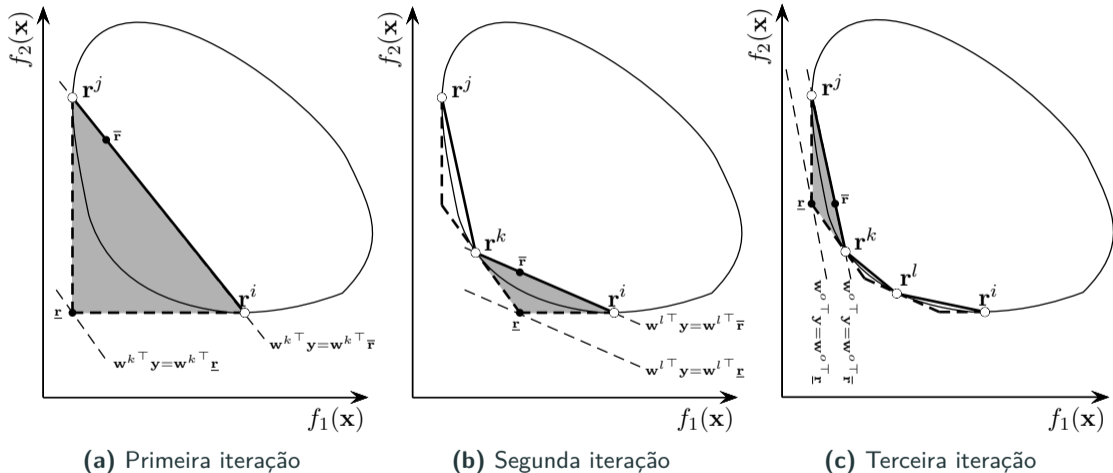


Figura 10: Sequência de passos ilustrativos do algoritmo MONISE.

Inicialização: Encontrar m soluções otimizando somente um objetivo em cada. Essas soluções formam uma vizinhança.

Loop:

1. Encontrar o hiperplano com maior distância entre a representação interna e externa da fronteira de Pareto considerando as soluções atuais.
2. Usando o vetor de parâmetros w da reta encontre outra solução \mathbf{r} usando a Definição 2.

Exemplo para a base de dados Monks

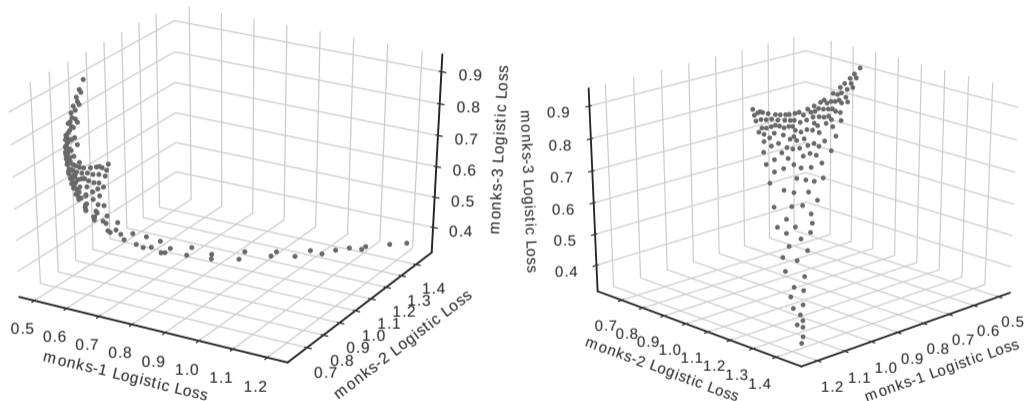


Figura 11: Perspectivas da representação da fronteira de Pareto, com o erro logístico de cada tarefa.

Modelos lineares e seus objetivos conflitantes

$$\min_{\zeta} \sum_{i=1}^L l_i(\zeta) + \sum_{j=1}^R \lambda_j r_j(\zeta) \quad (5)$$

onde $l_i(\zeta), i \in \{1, \dots, L\}$ são as perdas de aprendizado e $r_j(\zeta), j \in \{1, \dots, R\}$ são as regularizações.

E adaptando esse modelo para o problema de pesos temos:

$$\min_{\zeta} \sum_{i=1}^L w_i l_i(\zeta) + \sum_{j=1}^R w_{L+j} r_j(\zeta) \equiv w_1 f_1(\zeta) + \dots + w_m f_m(\zeta) \quad (6)$$

onde $f_1(\zeta), \dots, f_m(\zeta)$ são os m objetivos conflitantes do problema e w_1, \dots, w_m são os m vetores de pesos desses objetivos.

Filtragem e agregação de ensembles

- **Winner takes all** (wta)
- **Winner takes all por classe** (wtaPL)
- **Elite K** (elite)
- **Elite K por classe** (elitePL)
- **Filtragem multiobjetivo** (moPL)
- **Diversidade máxima** (max-div)

- **Voto simples** (svote)
- **Voto ponderado** (wv)
- **Soma de distribuições** (dsum)
- **Combinação Bayesiana** (bc)
- **Stacking** (stk)

Arcabouço proposto

1. **Modelagem multiobjetivo** - consiste em adaptar modelos de aprendizado de máquina para destacar os objetivos conflitantes.
2. **Treinamento multi-objetivo** - consiste em adotar métodos de otimização multiobjetivo, encontrando uma amostragem diversa da fronteira de Pareto.
3. **Seleção de modelo** ou **agregação de ensemble** - agrega múltiplas soluções eficientes para criar uma máquina de aprendizado.

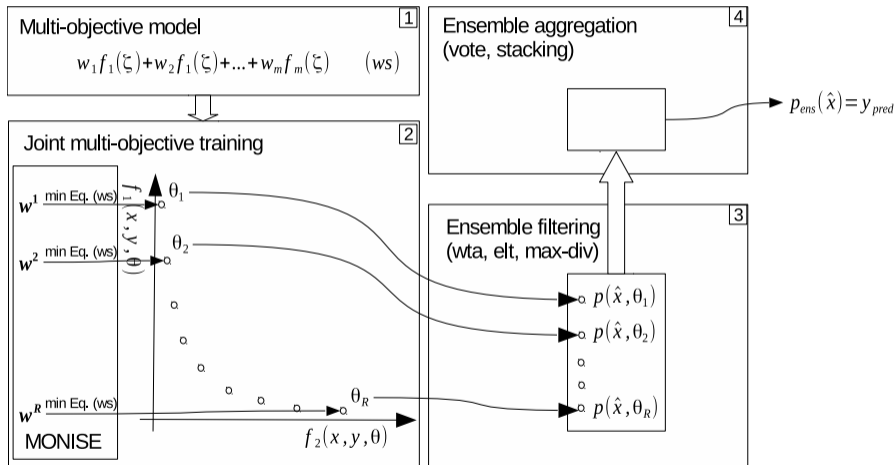


Figura 12: Visão geral do arcabouço proposto para aprendizado multiobjetivo.

Explorando conflitos em diferentes contextos

Classificação multi-classe

Algoritmos:

- kNN.
- Decision trees.
- Regressão logística e multinomial [Bishop, 2006].
- Máquinas de vetores-suporte [Cortes and Vapnik, 1995].
- Redes neurais artificiais.

Seleção de modelos:

- Conhecimento de especialista [Bergstra and Bengio, 2012, Chang and Lin, 2011].
- Busca em grid [Larochelle et al., 2007, Krstajic et al., 2014, Huang et al., 2012].
- Heurísticas [Kulaif and Von Zuben, 2013, Camilleri and Neri, 2014, Zheng and Bilenko, 2013].
- Estimação de distribuição [Weihs et al., 2005, Bergstra et al., 2011, Bergstra and Bengio, 2012].

Ensembles:

- Bagging [Breiman, 1996].
- Boosting [Schapire, 2009].
- Random forests [Breiman, 2001].

Conflito entre erro e complexidade em regressão multinomial [Raimundo et al., 2021]

Formulação original:

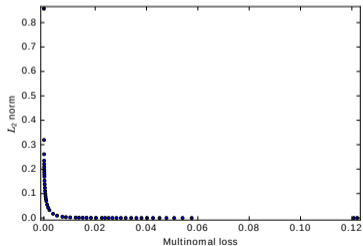
$$\min_{\theta} \sum_{k=1}^K \sum_{i=1}^N - \left[y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] + \lambda r(\theta),$$

Formulação adaptada:

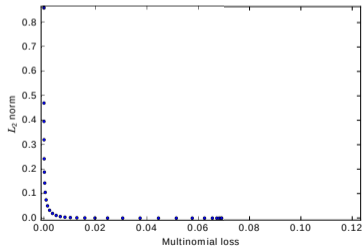
$$\min_{\theta} w_1 \sum_{k=1}^K \sum_{i=1}^N - \left[y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] + w_2 r(\theta),$$

Classificação multiclasse - seleção de modelos [Raimundo et al., 2021]

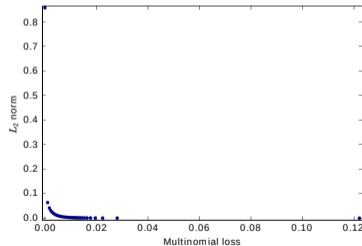
- 19 bases de dados (4 pastas).
- Modelo multinomial regularizado.
- Comparação com: busca em grade linear, busca em grade exponencial, Nelder-Mead, e hyperopt.
- Acurácia na validação.



(a) NISE



(b) Logarithmic grid search



(c) Constant grid search

Tabela 1: Comparação estatística envolvendo cinco técnicas de seleção de modelo variando o número de avaliações de função.

method	evals	rank	#<	#>
NISE	50	6.25	0	10
log grid	50	6.46	0	10
Hyperopt	50	6.71	0	10
NISE	25	6.73	0	10
log grid	25	6.97	0	9
const grid	50	7.70	4	6
Hyperopt	25	7.77	5	6
NISE	10	7.81	5	6
const grid	25	8.27	5	3
log grid	10	8.75	8	1
Nelder-Mead	50	8.80	8	1
Nelder-Mead	25	8.93	8	1
const grid	10	9.24	9	1
Hyperopt	10	9.45	9	0
Nelder-Mead	10	10.15	13	0

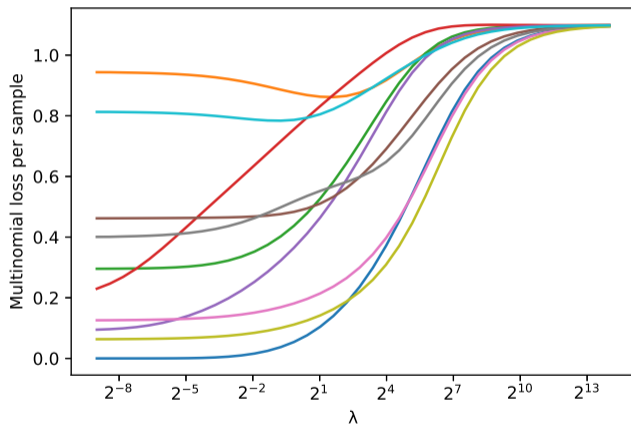
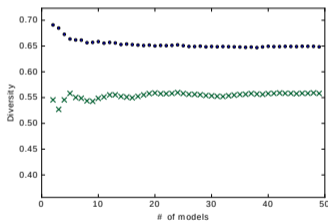


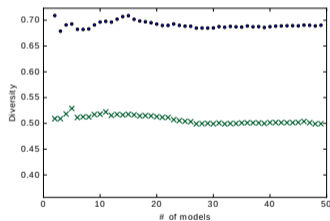
Figura 14: Ilustração da perda multinomial com o crescimento de λ para 10 amostras.

Classificação multiclasse - diversidade em ensembles

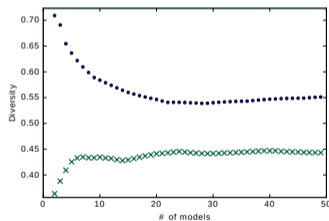
- 19 bases de dados (4 pastas).
- Modelo multinomial regularizado.
- Comparação de NISE e NISE ajustado com bagging e boosting.
- Avaliação da diversidade na validação.



(a) NISE



(b) Bagging



(c) Boosting

Figura 15: Evolução das métricas de diversidade (●) e both-correct (×) aumentando o número de componentes gerados para a base de dados *heart-Cleveland* diversity.

- 121 bases de dados (4 pastas).
- Modelo multinomial regularizado.
- Comparação de algumas formas de filtragem e agregação com 179 classificadores.
- Avaliação das métricas acurácia e kappa (dentre as pastas) no teste.

Tabela 2: Ranking de Friedman médio considerando acurácia.

method	rank	#<	#>	method	rank	#<	#>
rf_caret*	37.75	0	164	pnn_matlab*	63.78	7	110
parRF_caret*	38.03	0	162	cforest_caret*	63.79	7	110
svm_C*	40.69	0	161	gaussprRadial_R*	64.45	7	107
svmPoly_caret*	41.50	0	160	wta_svote	64.46	7	107
elm_kernel_matlab*	43.40	0	160	RandomForest_weka	65.14	7	107
svmRadialCost_caret*	43.47	0	160	svmLinear_caret*	66.56	12	102
rforest_R	44.38	0	160	dkp_C*	66.62	12	102
svmRadial_caret*	46.10	0	155	MultiBoostAB_RandomForest_weka	67.50	12	101
elite_svote	46.14	0	155	mlp_C*	68.97	16	98
elite_dsum	46.24	0	155	fda_caret	69.20	16	96
elitePL_dsum	47.13	0	155	RandomCommittee_weka	69.47	16	96
elitePL_svote	47.14	0	155	knn_caret*	69.49	16	96
max-div_svote	48.83	0	152	mlpWeightDecay_caret*	69.59	17	96
C5.0_caret*	48.93	0	152	Decorate_weka	69.78	18	95
avNNet_caret*	49.01	0	152	MultiBoostAB_MLP_weka	70.43	18	94
moPL_dsum	49.11	0	152	rda_R*	70.97	20	94
max-div_dsum	50.38	0	148	gcvEarth_caret	71.34	22	94
nnet_caret*	50.49	0	147	multinom_caret*	71.68	22	94
wtaPL_dsum	51.39	0	145	knn_R*	72.11	23	93
Bagging_LibSVM_weka*	51.58	0	145	MultiBoostAB_PART_weka	72.20	23	93
moPL_svote	51.99	0	144	glmnet_R	72.43	24	92
pcaNNet_caret*	52.12	0	144	trebag_caret	72.56	24	92
mlp_caret*	52.83	0	142	svmlight_C*	72.60	24	92
RotationForest_weka	53.29	0	140	mda_caret	72.66	24	92
wtaPL_svote	55.52	0	133	ClassificationViaRegression_weka	72.67	24	92
RRF_caret*	55.81	0	132	Bagging_PART_weka	73.12	24	91
MultiBoostAB_LibSVM_weka*	57.12	1	128	elm_matlab*	74.09	24	91
RRFglobal_caret*	57.17	1	128	SimpleLogistic_weka	74.75	25	89
LibSVM_weka*	58.29	2	126	pda_caret*	75.03	26	88
adaboost_R	60.42	3	119	rbfDDA_caret*	75.59	26	86

*Métodos com o ajuste de parâmetros feito com dados que podem estar no conjunto de teste, como reportado em Wainberg et al., 2016.

Tabela 3: Ranking de Friedman médio considerando kappa.

method	rank	#<	#>	method	rank	#<	#>
parRF_caret*	40.48	0	164	MultiBoostAB_PART_weka	63.74	3	111
rf_caret*	40.58	0	164	RandomCommittee_weka	66.47	4	104
svm_C*	43.88	0	161	MultiBoostAB_J48_weka	67.63	7	101
rforest_R	46.81	0	159	treebag_caret	67.64	7	101
elite_dsum	47.50	0	158	Bagging_PART_weka	67.80	7	101
mlp_caret*	47.78	0	158	LibSVM_weka*	67.82	7	101
elite_svote	48.27	0	158	MultiBoostAB_LibSVM_weka*	67.84	7	101
elitePL_dsum	49.07	0	151	RandomForest_weka	68.22	7	100
nnet_caret*	49.32	0	151	fda_caret	68.22	7	100
elitePL_svote	49.63	0	150	AdaBoostM1_J48_weka	68.69	9	97
elm_kernel_matlab*	50.13	0	149	rda_R*	69.05	10	97
C5.0_caret*	50.14	0	149	wta_svote	69.74	12	96
svmPoly_caret*	50.66	0	148	MultiBoostAB_RandomForest_weka	69.99	13	96
moPL_dsum	50.89	0	147	Bagging_RandomTree_weka	70.34	14	95
max_div_svote	51.73	0	145	mlp_C*	70.79	14	95
avNNet_caret*	51.79	0	145	gcvEarth_caret	71.02	16	95
svmRadialCost_caret*	52.28	0	144	multinom_caret*	72.10	19	95
RRF_caret*	52.29	0	144	MultilayerPerceptron_weka	72.75	19	92
wtaPL_dsum	52.92	0	144	Bagging_J48_weka	73.18	19	91
pcaNNet_caret*	54.08	0	140	svmLinear_caret*	73.19	19	91
moPL_svote	54.20	0	139	mda_caret	73.27	20	91
max_div_dsum	54.39	0	139	mlpWeightDecay_caret*	73.57	22	91
RRFglobal_caret*	54.60	0	138	gaussprRadial_R*	74.31	23	90
svmRadial_caret*	55.71	0	137	MultiBoostAB_RandomTree_weka	74.96	24	90
wtaPL_svote	57.67	0	128	pda_caret*	75.43	24	89
RotationForest_weka	59.18	0	126	pnn_matlab*	75.45	24	89
logitboost_R	62.13	2	115	ClassificationViaRegression_weka	75.50	24	88
adaboost_R	62.52	2	114	glmnet_R	75.66	24	88
MultiBoostAB_MLP_weka	62.73	2	114	SMO_weka	75.75	24	88
Decorate_weka	63.48	3	111	knn_caret*	75.81	24	88

*Métodos com o ajuste de parâmetros feito com dados que podem estar no conjunto de teste, como reportado em Wainberg et al., 2016.

Classificação multi-classe - classes desbalanceadas

Quando ocorre desbalanço de classe?

Quando existe uma grande diferença entre o número de amostras entre as classes.

É um problema encontrado naturalmente [SUN et al., 2009]:

- Detecção de fraude,
- Diagnóstico médico,
- Detecção de invasão em redes,
- Detecção de derramamento de óleo,
- Detecção de problemas de manufatura.

O desbalanço das classes pode deteriorar a performance de um classificador não especializado.

- **Cost-sensitive** pondera o erro de classificação de cada classe [Bradford et al., 1998, Lin et al., 2002, Datta and Das, 2015];
- **Reamostragem** sub-amostra ou sob-amostra classes [Chawla et al., 2002, He et al., 2008];
- **Boosting** adapta o AdaBoost [Sun et al., 2005, Chawla et al., 2003];
- **Otimização multi-objectivo** [Li et al., 2018, Soda, 2011].

O conflito entre classes [Raimundo and Von Zuben, 2020]

Regressão multinomial:

$$\min_{\theta} \sum_{k=1}^K \left[\sum_{i=1}^N -y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] \equiv \sum_{k=1}^K l_k(\theta), \quad (7)$$

$l_k(\theta)$, $k \in \{1, \dots, K\}$ podem ser conflitantes por que classificar corretamente amostras de uma classe pode induzir o erro de classificação de outras classes.

Com isso:

1. Agregação uniforme pode melhorar o erro de aprendizado focando na classe majoritária;
2. Além disso sem considerar o conflito, um usuário pode não ter sua preferência atendida: talvez focar no erro de aprendizado de uma classe seja mais importante que outro para uma certa situação.

O conflito entre classes [Raimundo and Von Zuben, 2020]

Formulação original:

$$\min_{\theta} \sum_{k=1}^K \frac{1}{u_k} \sum_{i=1}^N - \left[y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] + \lambda r(\theta).$$

Formulação adaptada I:

$$\min_{\theta} w_1 \sum_{k=1}^K \frac{1}{u_k} \sum_{i=1}^N - \left[y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] + w_2 r(\theta).$$

Formulação adaptada II:

$$\min_{\theta} \sum_{k=1}^K w_k \left[-\frac{1}{u_k} \sum_{i=1}^N y_i^k \ln \left(\frac{e^{\theta_k^\top \phi(\mathbf{x}_i)}}{\sum_{j=1}^K e^{\theta_j^\top \phi(\mathbf{x}_i)}} \right) \right] + w_{K+1} r(\theta).$$

O conflito entre classes [Raimundo and Von Zuben, 2020]

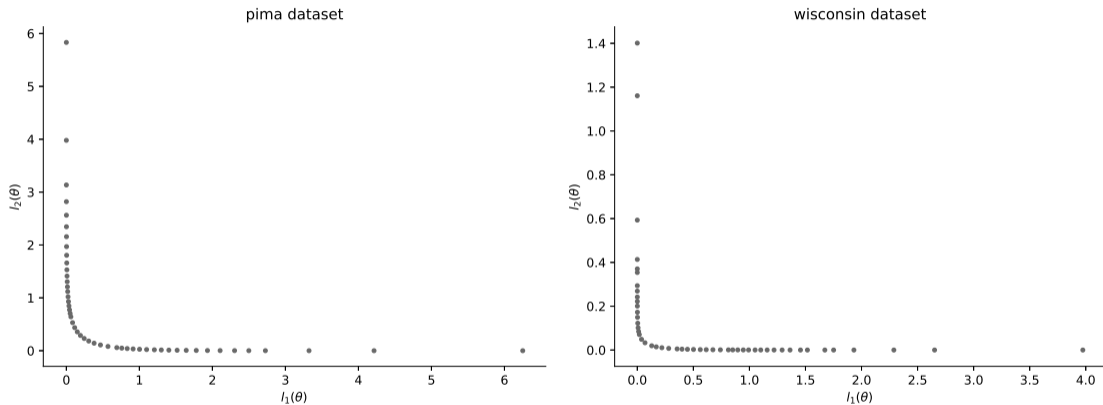


Figura 16: Representações da fronteira de Pareto associadas às bases de dados Pima e Wisconsin (Available at archive.ics.uci.edu/ml)

Abordagens propostas [Raimundo and Von Zuben, 2020]

O arcabouço proposto consiste em explorar diferentes combinações de modelos gerados por otimização multi-objetivo.

- **MO**: 150 modelos que exploram o conflito entre as classes;
- **StandardMO**: 50 modelos com preferência constante ($u_k = 1$);
- **CSMO** 50 modelos com preferência inversamente proporcional ao número de amostras da classe ($u_k = n_k$);
- **MO&AllAdHoc** combina as três propostas (**MO**, **StandardMO**, e **CSMO**);
- **AllAdHoc** combina as propostas adhoc **StandartMO** e **CSMO**.

Resultados [Raimundo and Von Zuben, 2020]

Tabela 4: Ranking médio de Friedman rank para g-mean, kappa e F_1 nas bases de dados com desbalanço maior que 9 – KEEL dataset.

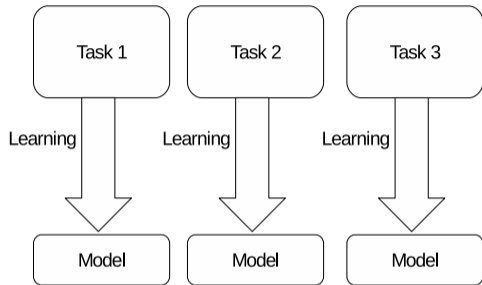
Method	kappa			F_1			g-mean			non-dom		
	Rank	#<	#>	Rank	#<	#>	Rank	#<	#>	Rank	#<	#>
MO&AllAdHoc	8.12	0	3	7.01	0	7	7.10	0	10	5.18	0	13
AllAdHoc	8.46	0	3	9.45	0	1	6.84	0	10	6.29	0	12
CSMO	9.99	1	1	10.90	3	1	5.09	0	13	6.69	0	11
MO	8.63	0	3	8.96	0	1	7.77	1	6	6.79	0	11
SMOTEBoost	7.32	0	6	7.22	0	6	7.88	1	6	7.23	0	10
RAMOBoost	7.81	0	3	7.83	0	5	7.94	1	5	8.05	1	6
SMOTEENN	8.72	0	3	8.41	0	3	9.85	5	4	8.75	2	3
SMOTETomek	8.83	0	3	8.56	0	2	10.11	5	4	9.55	4	2
EasyEnsemble	14.32	16	0	15.16	17	0	6.98	0	10	9.85	5	2
CSManual	10.31	1	1	10.83	3	1	5.94	0	10	9.87	5	2
RndUnderSamp	8.44	0	3	8.34	0	3	10.25	7	3	10.13	5	2
RndOverSamp	8.99	0	3	8.83	0	1	9.82	5	4	10.39	5	1
SMOTE	8.73	0	3	8.60	0	2	9.78	5	4	10.89	6	1
StandardMO	9.88	0	1	9.91	2	1	13.53	14	0	10.93	6	1
ADASYN	8.72	0	3	8.63	0	2	10.49	8	3	10.97	6	1
ENN	10.06	1	1	9.77	1	1	12.51	12	1	11.94	7	1
TomekLinks	11.67	11	1	11.12	5	1	13.74	14	0	12.56	11	0
StandardManual	11.91	11	0	11.37	8	1	15.25	15	0	14.83	16	0

Tabela 5: Ranking médio de Friedman rank para g-mean, kappa e F_1 nas bases de dados com desbalanço maior que 9 – KEEL dataset.

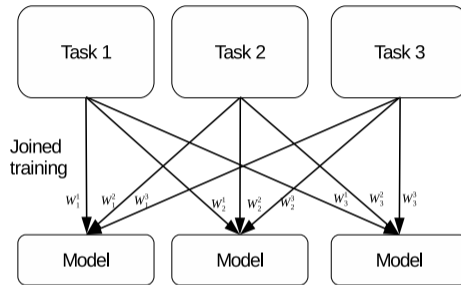
Method	kappa			F_1			g-mean			non-dom		
	Rank	#<	#>	Rank	#<	#>	Rank	#<	#>	Rank	#<	#>
StkMO&AllAdHoc	3.06	0	1	3.12	0	1	2.92	-	-	2.42	0	4
EltMO&AllAdHoc	3.10	0	1	3.14	0	1	3.35	-	-	3.07	0	2
MO&AllAdHoc	3.28	0	1	3.02	0	1	3.51	-	-	3.44	1	1
SMOTEBoost	3.02	0	1	3.04	0	1	3.79	-	-	3.63	1	0
RAMOBoost	3.24	0	1	3.20	0	1	3.92	-	-	4.10	2	0
EasyEnsemble	5.27	5	0	5.44	5	0	3.48	-	-	4.32	3	0

Aprendizado multitarefa

Compartilhamento de aprendizado



(a) Aprendizado mono-tarefa



(b) Aprendizado multi-tarefa

Formulação geral em aprendizado multi-tarefa.

$$\min_{\theta} \sum_{t=1}^T l(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \theta^{(t)}) + \lambda r(\theta).$$

Compartilhamento de aprendizado

Classificação multirrótulo:

- Binary Relevance (BR) [Zhang and Zhou, 2007].
- Label powerset (LP) [Read et al., 2008].
- Classification Chains (CC) [Read et al., 2011, Ramírez-Corona et al., 2016, Dembczy, 2010].
- Aprendizado multitarefa [Gonçalves et al., 2015, Japkowicz and Matwin, 2015].

Aprendizado multitarefa:

- Redes neurais (feature learning) [Caruana, 1998].
- Através de regularização [Kim and Paik, 2014, Obozinski et al., 2008].
- Através de ensembles [Simm et al., 2014, Faddoul et al., 2012].

Também podemos incluir nessa classe o aprendizado por transferência.

Formulação geral original:

$$\min_{\theta} \sum_{t=1}^T l(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \theta^{(t)}) + \lambda r(\theta).$$

Formulação adaptada:

$$\min_{\theta} \sum_{t=1}^T \mathbf{w}_t l(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \theta) + \mathbf{w}_{T+1} \|\theta\|.$$

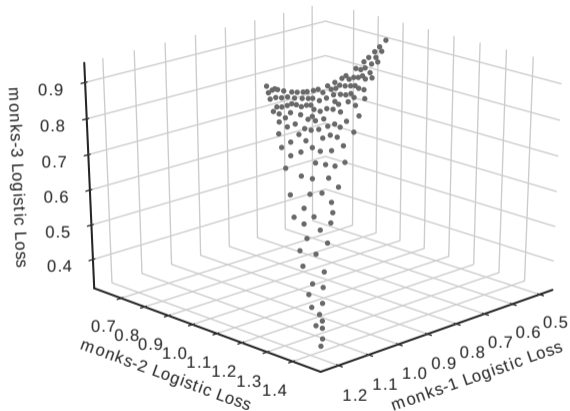


Figura 17: Fronteira de Pareto para a base de dados monks.

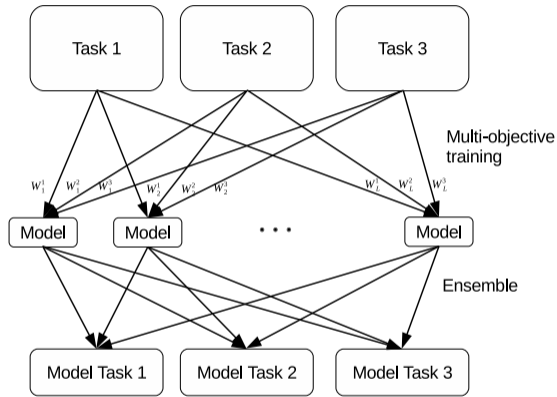
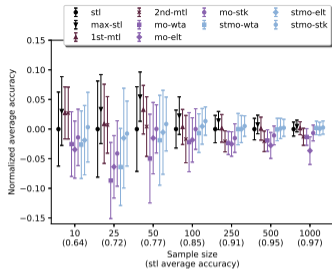


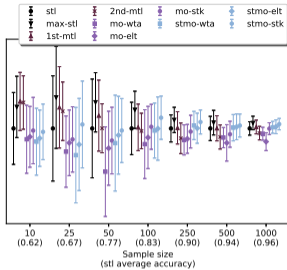
Figura 18: Representação da metodologia multi-objetivo para aprendizado multi-tarefa.

Aprendizado multitarefa [Raimundo and Von Zuben, 2018a]

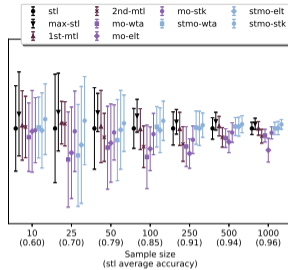
- 9 bases de dados artificiais (3 perfis de transferência de aprendizado com 3 tipos de ruído), 3 bases de dados reais.
- Modelo de aprendizado multitarefa com somente um vetor de parâmetros.
- 6 métodos propostos com aprendizado multiobjetivo multitarefa seguido por três formas de agregação, e com aprendizado MOO+MTL e modelos STL para auxiliar nas tarefas discrepantes, com três formas de agregação.
- Comparação em termos de acurácia aumentando o número de amostras



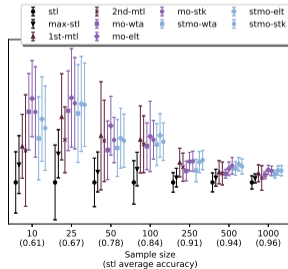
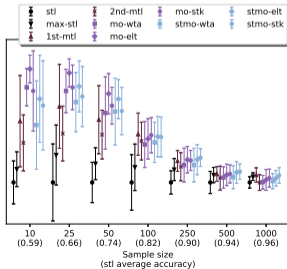
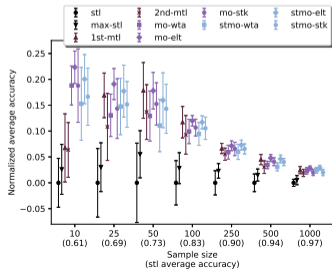
(a) independent

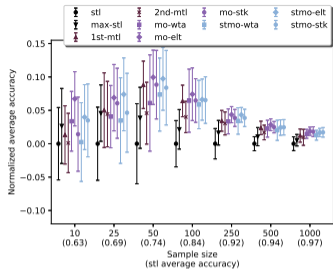


(b) + dirty

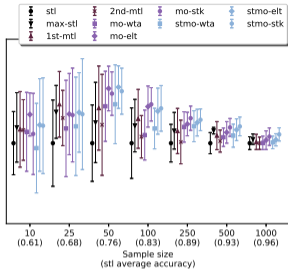


(c) + outliers

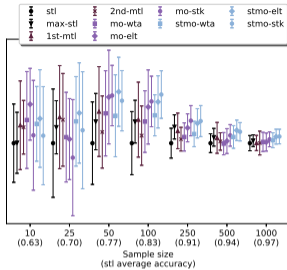




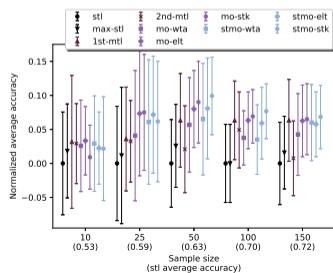
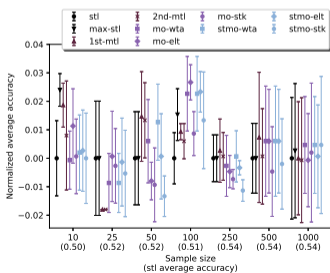
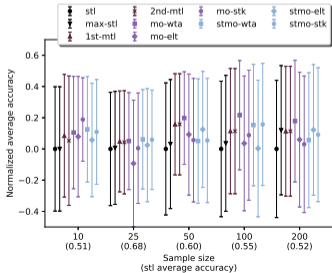
(a) three clusters



(b) + dirty



(c) + outliers



Aprendizado multitarefa - Relação entre as tarefas

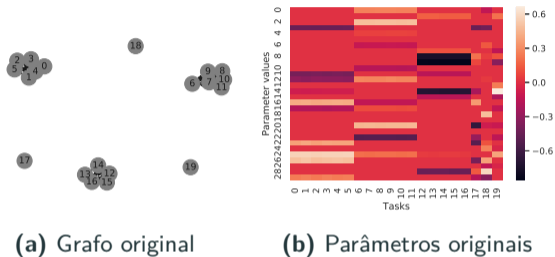


Figura 21: Representação das relações entre as tarefas para a base de dados com três clusters e outliers.

Influência média:

$$\mathbf{u}^t = \left[\frac{1}{10} \sum_{i \in \mathcal{B}^t} \mathbf{w}_1^{(i)}, \dots, \frac{1}{10} \sum_{i \in \mathcal{B}^t} \mathbf{w}_T^{(i)} \right]^\top$$

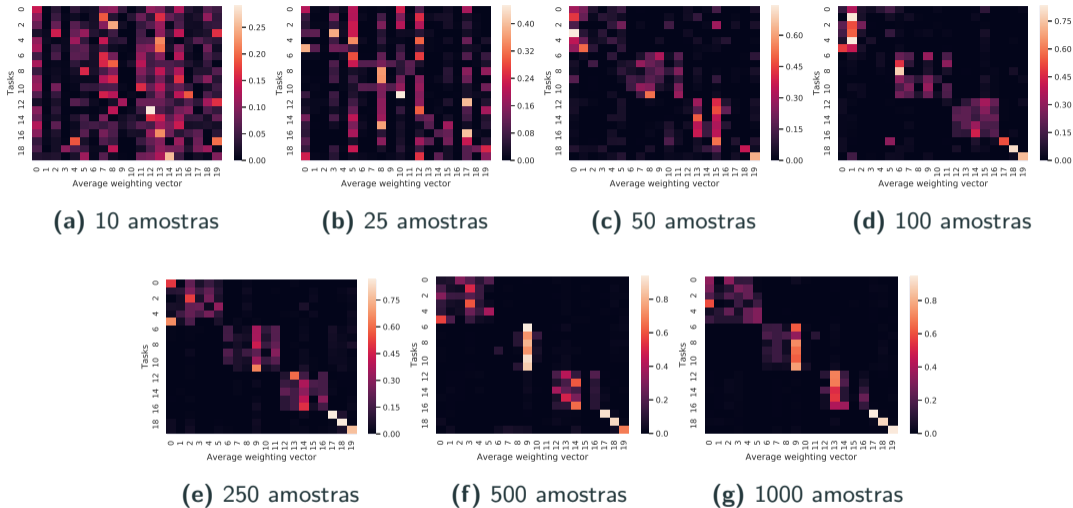


Figura 22: Representação da influência média dos modelos multitarefa treinados sob a perspectiva multiobjetivo para a base de dados com três clusters e outliers.

Outras contribuições em transferência de aprendizado

- Detecção de crises epilépticas usando aprendizado por transferência com múltiplas extrações de dados [Beserra et al., 2018, Beserra et al., 2017].
- Classificação multirrótulo [Raimundo and Von Zuben, 2018b].

Imparcialidade em aprendizado de máquina

O problema da discriminação em aprendizado de máquina

- Modelo que discriminava mulheres na avaliação de currículos [Dastin, 2018].
- Modelo que discriminava Afro-Americanos dando um risco maior de reincidência em crime mesmo com o mesmo perfil de um branco [Julia Angwin and Kirchner, 2016, Dieterich et al., 2016, Dressel and Farid, 2018].

Se a base de dados contém discriminação, um modelo de aprendizado pode propagar essa discriminação caso não leve isso em conta.

Considere uma base de dados $\mathcal{D} = \{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^N$ sendo $\mathbf{x}_i \in \mathcal{X}$ a entrada $y_i \in \mathcal{Y}$ a saída, e $a_i \in \mathcal{A}$ é um atributo protegido (e.g., gênero ou etnia). Sendo \mathcal{D}^a os indivíduos de cada grupo do atributo protegido $a \in \mathcal{A}$, o modelo f_θ é considerado justo para uma função de benefício $\delta(f_\theta)$ se os valores forem similares para todo $a \in \mathcal{A}$.

Conflito entre funções de benefício

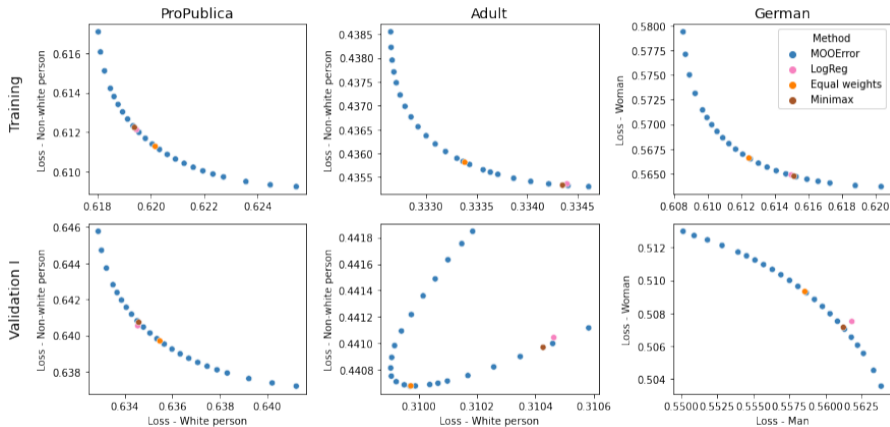
Conflito entre as perdas de aprendizado:

$$\underset{\theta}{\text{minimize}} \quad \sum_{j=1}^{|\mathcal{A}|} w_j \cdot \ell^j(\theta) \quad (8)$$

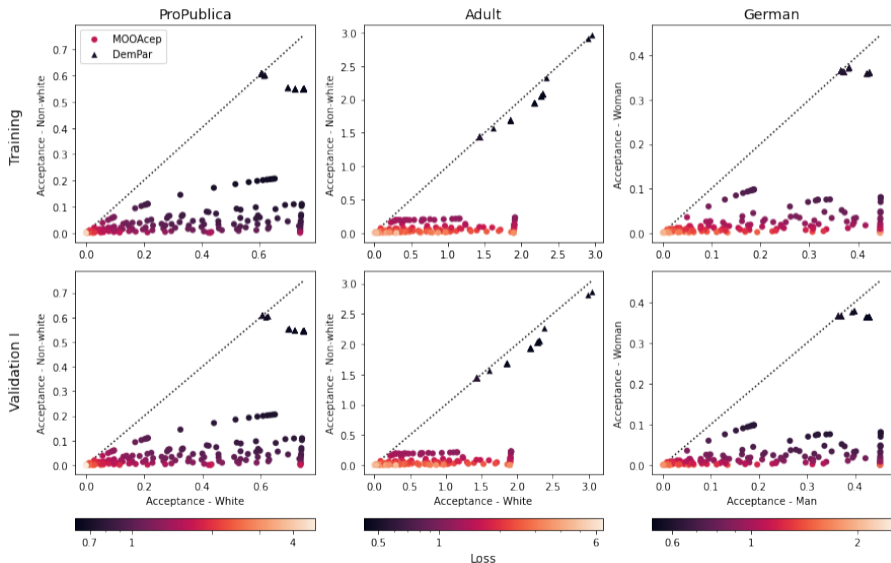
Conflito entre as aceitações (erro de aprendizado para alcançar uma saída desejada):

$$\underset{\theta}{\text{minimize}} \quad \sum_{j=1}^{|\mathcal{A}|} w_j \cdot \alpha^j(\theta) + w_{|\mathcal{A}|+1} \cdot l(\theta) \quad (9)$$

Parcialidade em erro de aprendizado



Parcialidade em aceitação/atingir a saída desejável



Outros trabalhos

Múltiplas visões em detecção taxonômica de formigas [Marques et al., 2018]

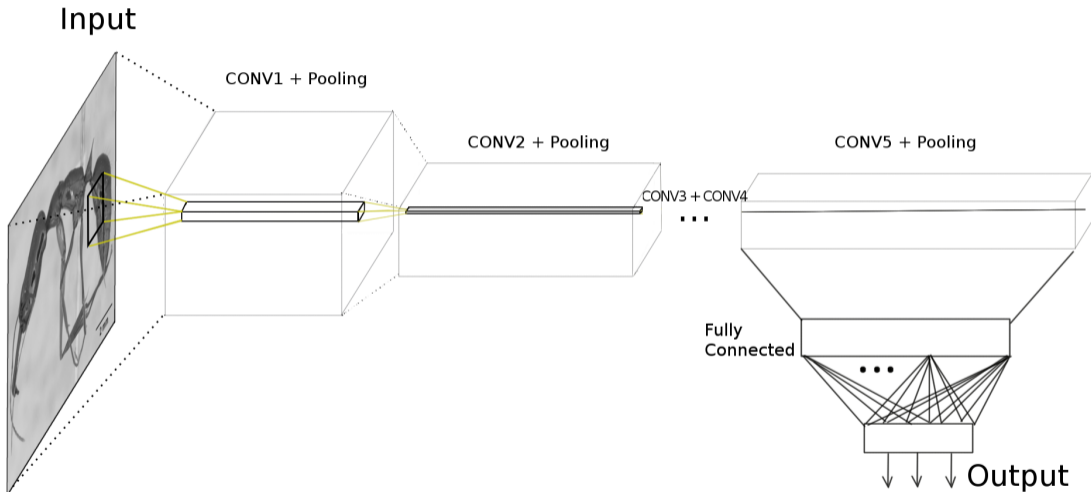


Figura 23: Modelo ResNet para classificação de gênero taxonômico de formiga.

Múltiplas visões em detecção taxonômica de formigas [Marques et al., 2018]

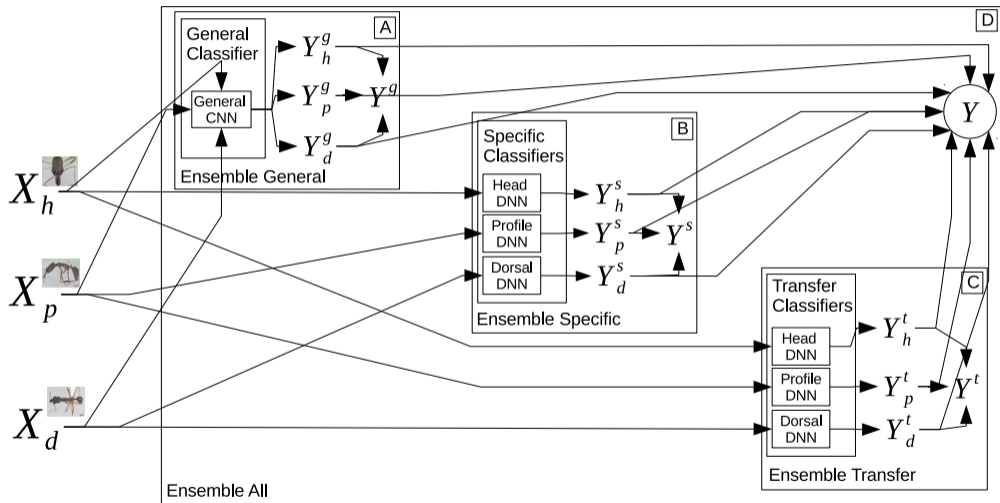


Figura 24: Arcabouço para lidar com múltiplas visões do mesmo espécime.

Group LASSO e o conflito em múltiplas visões

Considerando V visões, e seja \mathcal{V}^v o conjunto que contém os índices do grupo v . A formulação Group LASSO [Yuan and Lin, 2006] é dada por:

$$\min_{\theta} l(\mathbf{x}, \mathbf{y}, \theta) + \lambda \sum_{v=1}^V \sqrt{|\mathcal{V}^v|} \sqrt{\sum_{i \in \mathcal{V}^v} \theta_i^2}, \quad (10)$$

Agora considerando que a perda de aprendizado é conflitante com a regularização em cada grupo, é possível modelar o problema com otimização multi-objetivo [Raimundo and Von Zuben, 2019]:

$$\min_{\theta} w^l l(\mathbf{x}, \mathbf{y}, \theta) + \sum_{v=1}^V w_v^r \sqrt{\sum_{i \in \mathcal{V}^v} \theta_i^2}. \quad (11)$$

Múltiplas visões em detecção de crises epiléticas [Raimundo and Von Zuben, 2019]

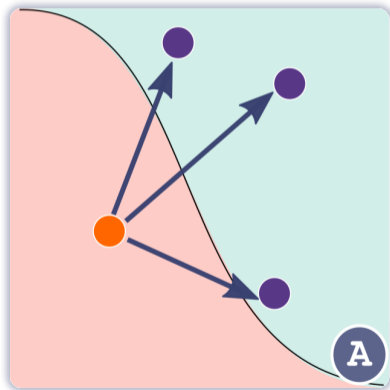
Tabela 6: Valores médios para as métricas SEN, SPE e LAT.

	fou			gph			wlt			gl			mp		
	wta	elt	stk	wta	elt	stk	wta	elt	stk	wta	elt	stk	wta	elt	stk
SEN	0.833	0.826	0.853	0.783	0.775	0.735	0.745	0.771	0.789	0.829	0.822	0.836	0.831	0.839	0.862
SPE	0.953	0.951	0.935	0.857	0.863	0.886	0.924	0.934	0.916	0.953	0.954	0.943	0.959	0.960	0.921
LAT	2.634	2.810	2.029	3.130	3.739	3.995	2.432	2.302	2.394	3.088	3.117	2.117	2.697	2.373	2.180

- **SEN** - mede a proporção de segundo com crises epiléticas que foram corretamente identificadas.
- **SPE** - mede a proporção de segundo com sem crises epiléticas que foram corretamente identificadas.
- **LAT** - número de segundos para identificar uma crise.

Antecedentes contrafactuais – o que é?

"Se a sua concentração glicose plasmática fosse 158.3 e seu nível de insulina em 2 horas fosse 160.5, seu risco de diabetes cairia para 0.51."



	Features	Original Values	CFA Values
CF1	Passengers	4.11	1.33
CF2	Expansion Phase	1	5
CF3	Bus Stops HighIncomeHolder	0.69 0.0	0.47 0.15
CF4	Bus Stos HighIncomeHolder	0.69 0.0	0.41 0.07

Figura 25: Exemplo do conceito de antecedentes contrafactuais.

Antecedentes contrafactuais – Crime em São Paulo

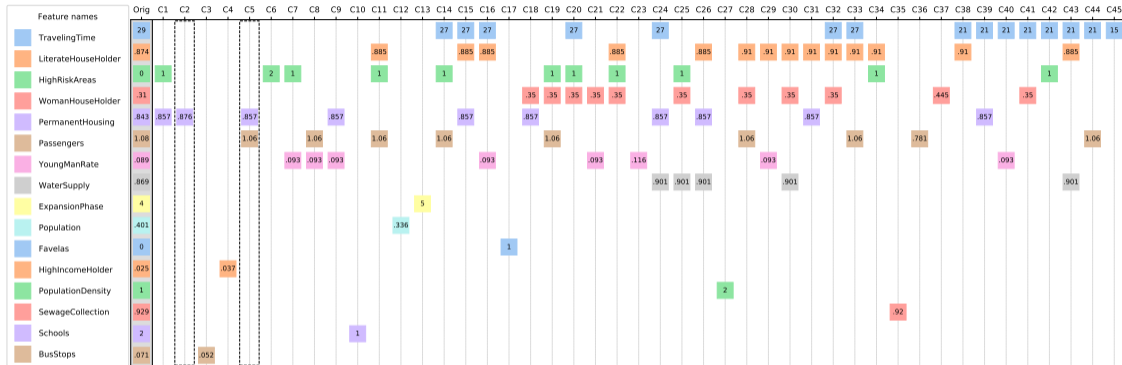


Figura 26: Enumeração de antecedentes contrafactuais Pareto-ótimos considerado cada atributo como uma função objetivo. Coluna Orig é a amostra original e as colunas de C1 a C28 são os antecedentes contrafactuais.

Antecedentes contrafactuais – Análise de crédito

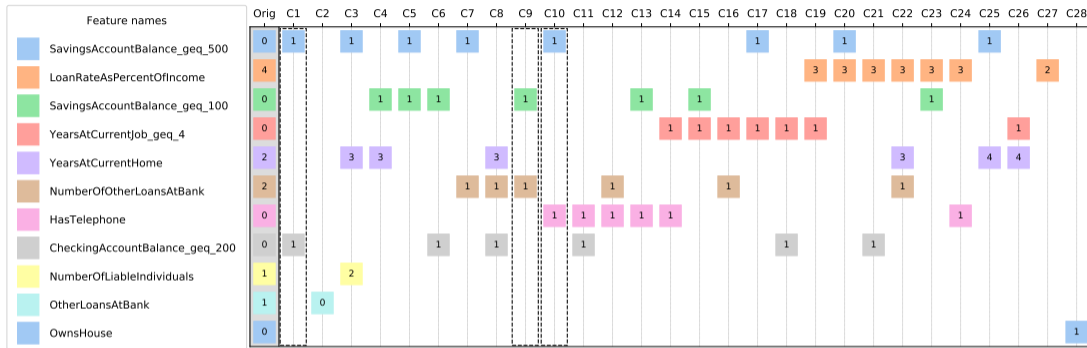


Figura 27: Enumeração de antecedentes contrafactuais Pareto-ótimos considerado cada atributo como uma função objetivo. Coluna Orig é a amostra original e as colunas de C1 a C28 são os antecedentes contrafactuais.




Conclusão




Aprendizado Multi-objetivo

Usando conflitos para aprender

Marcos M. Raimundo

3 de Setembro de 2021 - Campinas - Brazil

-  Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011).
Algorithms for Hyper-Parameter Optimization.
Advances in Neural Information Processing Systems (NIPS), pages 2546–2554.
-  Bergstra, J. and Bengio, Y. (2012).
Random Search for Hyper-Parameter Optimization.
Journal of Machine Learning Research, 13:281–305.
-  Beserra, F. S., Raimundo, M. M., and B, F. J. V. Z. (2017).
Multi-objective transfer learning for epileptic seizure detection.
In *Journal of Epilepsy and Clinical Neurophysiology*, volume 2, page 67.

-  Beserra, F. S., Raimundo, M. M., and B, F. J. V. Z. (2018).
Ensembles of Multiobjective-Based Classifiers for Detection of Epileptic Seizures.
In *Lecture Notes in Computer Science, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2017.*, volume 10657, pages 575–583.
-  Bishop, C. M. (2006).
Pattern Recognition and Machine Learning.
Springer.
-  Bradford, J., Kunz, C., Kohavi, R., Brunk, C., and Brodley, C. (1998).
Pruning decision trees with misclassification costs.
Machine Learning: ECML- 98, 1398:131 – 136.



Breiman, L. (1996).

Bagging predictors.

Machine Learning, 24(2):123–140.



Breiman, L. (2001).

Random forests.

Machine learning, 45:5–32.



Camilleri, M. and Neri, F. (2014).

An Algorithmic Approach to Parameter Selection in Machine Learning using Meta-Optimization Techniques.

WSEAS Transactions on Systems, 13:203–212.



Caruana, R. (1998).

A Dozen Tricks with Multitask Learning.

In *Neural Networks: Tricks of the Trade*, pages 165–191.



Chang, C.-c. and Lin, C.-j. (2011).

LIBSVM : A Library for Support Vector Machines.





ACM Transactions on Intelligent Systems and Technology (TIST), 2:1–39.



Chawla, N. V., Bowyer, K. W., and Hall, L. O. (2002).

SMOTE : Synthetic Minority Over-sampling Technique.

Artificial Intelligence, 16:321–357.

-  Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003).
SMOTEBoost : Improving Prediction.
Lecture Notes in Computer Science, 2838:107–119.
-  Cohon, J. L., Church, R. L., and Sheer, D. P. (1979).
Generating multiobjective trade-offs: An algorithm for bicriterion problems.
Water Resources Research, 15(5):1001–1010.
-  Cortes, C. and Vapnik, V. (1995).
Support-Vector Networks.
Machine Learning, 20(3):273–297.
-  Dastin, J. (2018).
Amazon scraps secret ai recruiting tool that showed bias against women.



Datta, S. and Das, S. (2015).

Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs.

Neural Networks, 70:39–52.



Dembczy, K. (2010).

Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains.




Proceedings of the 27th International Conference on Machine Learning, pages 279–286.







Dieterich, W., Mendoza, C., and Brennan, T. (2016).

Compas risk scales: Demonstrating accuracy equity and predictive parity.

Northpoint Inc, 7(7.4):1.

-  Dressel, J. and Farid, H. (2018).
The accuracy, fairness, and limits of predicting recidivism.
Science advances, 4(1):eaao5580.
-  Faddoul, J. B., Chidlovskii, B., Gilleron, R., and Torre, F. (2012).
Learning multiple tasks with boosted decision trees.
In *Machine Learning and Knowledge Discovery in Databases*, volume 7523, pages 681–696. Springer Berlin Heidelberg.
-  Gonçalves, A. R., Zuben, F. J. V., and Banerjee, A. (2015).
Multi-Label Structure Learning with Ising Model Selection.
In *Proceedings of 24th International Joint Conference on Artificial Intelligence*, pages 3525–3531.

-  He, H., Bai, Y., Garcia, E. A., and Li, S. (2008).
Adaptive Synthetic Sampling Approach for Imbalanced Learning.
(3):1322–1328.
-  Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2012).
Extreme learning machine for regression and multiclass classification.
IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics, 42(2):513–29.
-  Japkowicz, N. and Matwin, S. (2015).
Multi-label Classification via Multi-target Regression on Data Streams.
In *Lecture Notes in Computer Science*, volume 9356, pages 170–185.
-  Julia Angwin, Jeff Larson, S. M. and Kirchner, L. (2016).
Machine bias.



Kim, H. and Paik, J. (2014).

Low-Rank Representation-Based Object Tracking Using Multitask Feature Learning with Joint Sparsity.

Abstract and Applied Analysis, 2014:1–12.



Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014).

Cross-validation pitfalls when selecting and assessing regression and classification models.




Journal of Cheminformatics, 6(1):1–15.






Kulaif, A. C. P. and Von Zuben, F. J. (2013).

Improved Regularization in Extreme Learning Machines.

Anais do Congresso Brasileiro de Inteligência Computacional (CBIC).

-  Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007).
An empirical evaluation of deep architectures on problems with many factors of variation.
International Conference on Machine Learning, (2006):473–480.
-  Li, J., Fong, S., Wong, R. K., and Chu, V. W. (2018).
Adaptive multi-objective swarm fusion for imbalanced data classification.
Information Fusion, 39:1–24.
-  Lin, Y., Lee, Y., and Wahba, G. (2002).
Support vector machines for classification in nonstandard situations.
Machine Learning, 46(1-3):191–202.

-  Marques, A. C. R., M. Raimundo, M., B. Cavalheiro, E. M., FP Salles, L., Lyra, C., and J. Von Zuben, F. (2018).
Ant genera identification using an ensemble of convolutional neural networks.
PLoS One, 13(1):e0192011.
-  Obozinski, G., Wainwright, M. J., and Jordan, M. L. (2008).
High-dimensional support union recovery in multivariate regression.
In *Advances in Neural Information Processing Systems*, pages 1217–1224.
-  Raimundo, M. M., Drumond, T. F., Marques, A. C. R., Lyra, C., Rocha, A., and Von Zuben, F. J. (2021).
Exploring multiobjective training in multiclass classification.
Neurocomputing, 435:307–320.



Raimundo, M. M. et al. (2018).

Multi-objective optimization in machine learning: Otimização multiobjetivo em aprendizado de máquina.



Raimundo, M. M., Ferreira, P. A., and Von Zuben, F. J. (2020).

An extension of the non-inferior set estimation algorithm for many objectives.
European Journal of Operational Research, 284(1):53–66.



Raimundo, M. M. and Von Zuben, F. J. (2017).

MONISE - Many Objective Non-Inferior Set Estimation.
arXiv, 1709.00797:1–39.



Raimundo, M. M. and Von Zuben, F. J. (2018a).

Investigating multiobjective methods in multitask classification.

In *International Joint Conference on Neural Networks (IJCNN)*.



Raimundo, M. M. and Von Zuben, F. J. (2018b).

Many-Objective Ensemble-Based Multilabel Classification.




In Mendoza, M. and Velastín, S., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 365–373, Cham. Springer International Publishing.



Raimundo, M. M. and Von Zuben, F. J. (2019).

Investigating multiobjective methods in multitask classification.

In *Journal of Epilepsy and Clinical Neurophysiology*, page 14.

-  Raimundo, M. M. and Von Zuben, F. J. (2020).
Multi-criteria analysis involving pareto-optimal misclassification tradeoffs on imbalanced datasets.
In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.
-  Ramírez-Corona, M., Sucar, L. E., and Morales, E. F. (2016).
Hierarchical multilabel classification based on path evaluation.
International Journal of Approximate Reasoning, 68:179–193.
-  Read, J., Pfahringer, B., and Holmes, G. (2008).
Multi-label classification using ensembles of pruned sets.
Proceedings - IEEE International Conference on Data Mining, ICDM, pages 995–1000.

 Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011).

Classifier chains for multi-label classification.

Machine Learning, 85(3):333–359.

 Schapire, R. E. (2009).

A Short Introduction to Boosting.

Journal of Japanese Society for Artificial Intelligence, 14(5):771–780.

 Simm, J., Magrans, I., Abril, D. E., and Sugiyama, M. (2014).

Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression.

In *IEICE Transactions on Information and Systems*, number 6, pages 1677–1681.



Soda, P. (2011).

A multi-objective optimisation approach for class imbalance learning.

Pattern Recognition, 44(8):1801–1810.



Sun, Y., Wong, A., and Wang, Y. (2005).

Parameter inference of cost-sensitive boosting algorithms.





Machine Learning and Data Mining in Pattern Recognition, 3587(July):21–30.



SUN, Y., WONG, A. K. C., and KAMEL, M. S. (2009).

CLASSIFICATION OF IMBALANCED DATA: A REVIEW.

International Journal of Pattern Recognition and Artificial Intelligence, 23(04):687–719.

-  Wainberg, M., Alipanahi, B., and Frey, B. J. (2016).
Are Random Forests Truly the Best Classifiers?
Journal of Machine Learning Research, 17(110):1–5.
-  Weihs, C., Luebke, K., and Czogiel, I. (2005).
Response Surface Methodology for Optimizing Hyper Parameters.
-  Yuan, M. and Lin, Y. (2006).
Model selection and estimation in regression with grouped variables.
Journal of the Royal Statistical Society. Series B: Statistical Methodology, 68(1):49–67.
-  Zhang, M. L. and Zhou, Z. H. (2007).
ML-KNN: A lazy learning approach to multi-label learning.
Pattern Recognition, 40(7):2038–2048.



Zheng, A. X. and Bilenko, M. (2013).

Lazy Paired Hyper-Parameter Tuning.

International joint conference on Artificial Intelligence, pages 1924–1931.