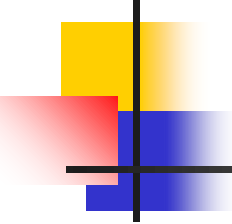


PREVISÃO DA QUANTIDADE DE CLASSES EM CLASSIFICAÇÃO HIERÁRQUICA MULTIRRÓTULO

Thissiany Beatriz Almeida
Helyane Bronoski Borges

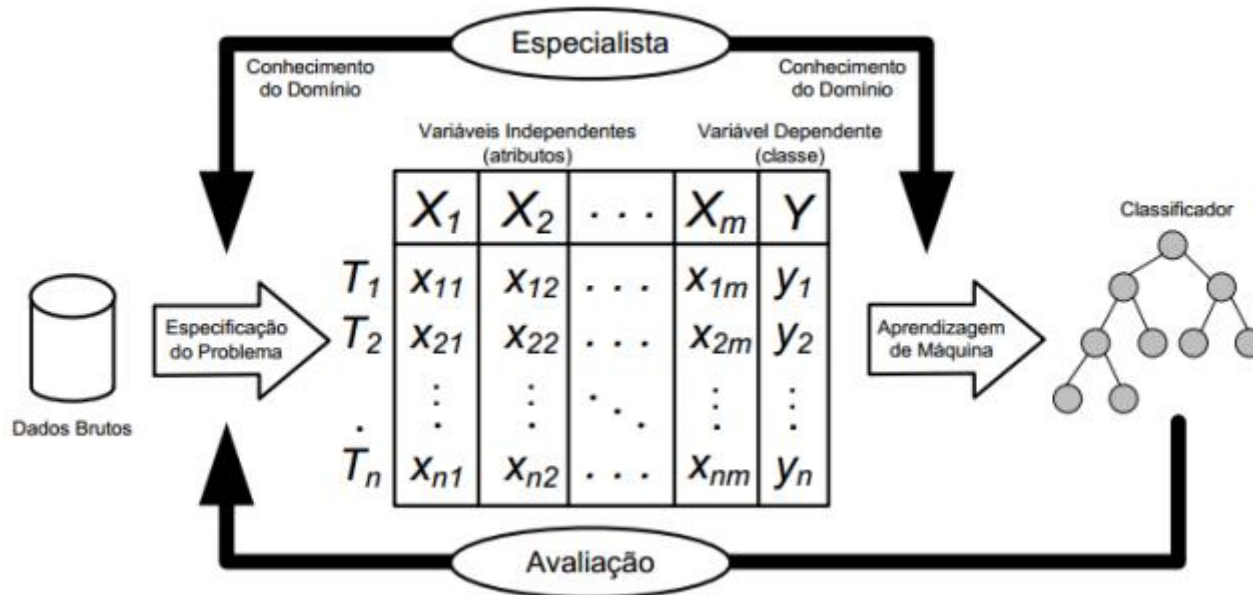


Agenda

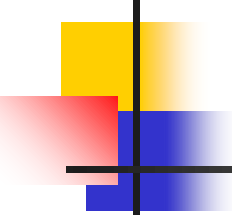
- Introdução
- Conceitos Fundamentais
- Metodologia
- Experimentos e Resultados
- Conclusão

Introdução

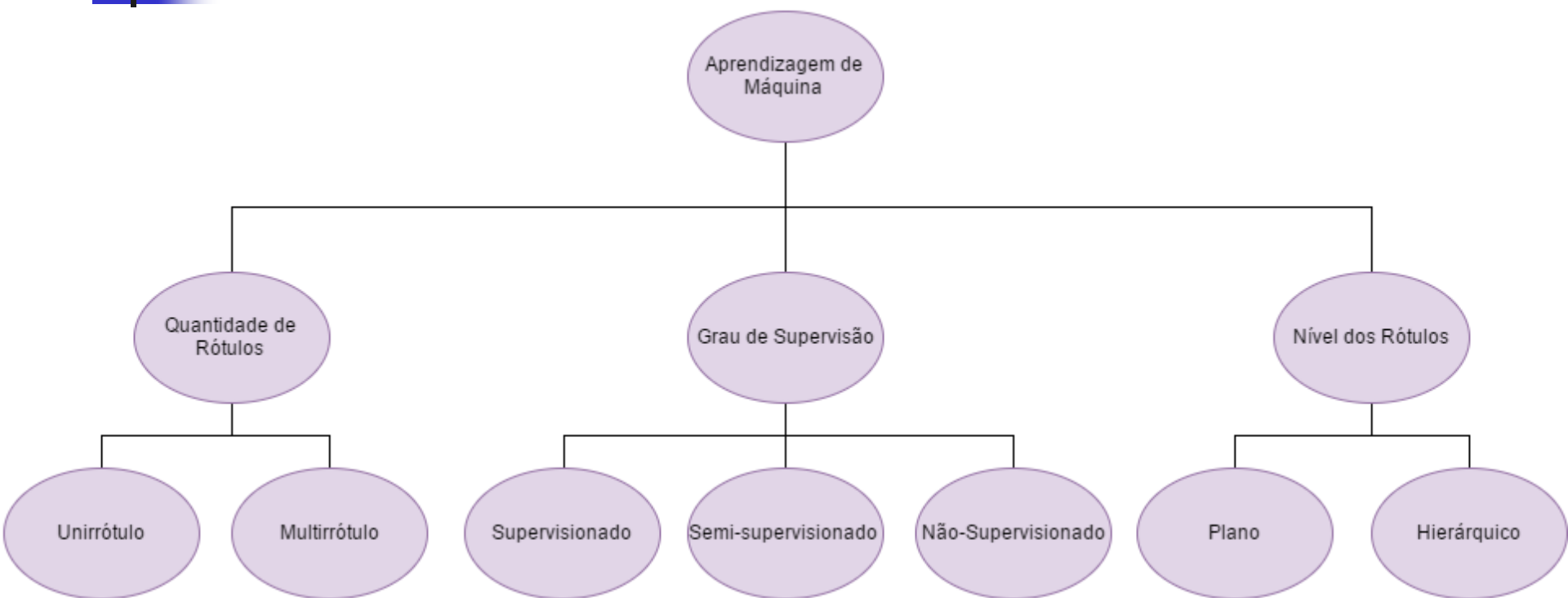
■ Classificação



Processo de Classificação.
Fonte: REZENDE, 2005.

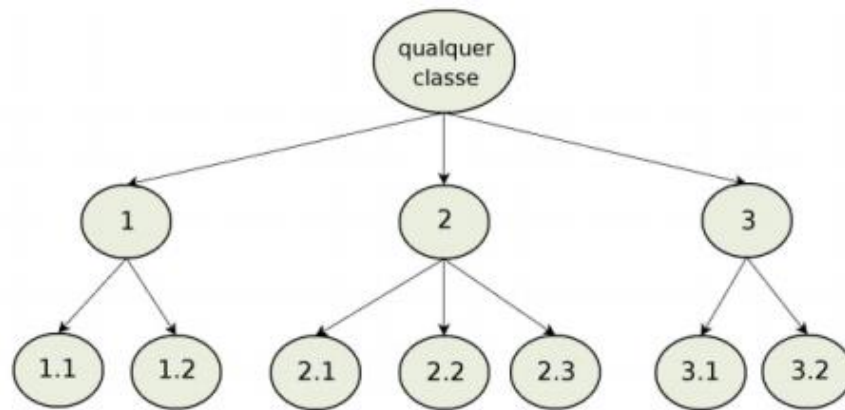


Introdução

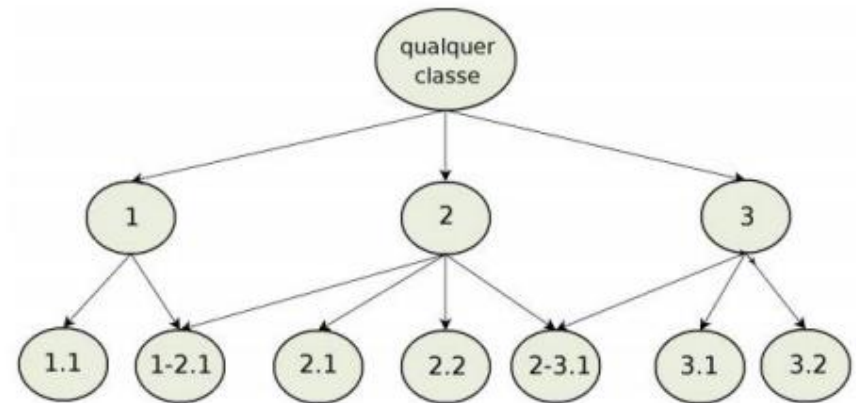


Introdução

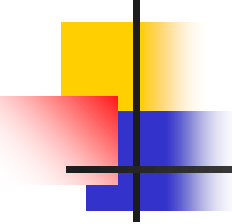
■ Classificação Hierárquica



Hierarquia de Classes estruturada como árvore.
Fonte: Carvalho et al, 2007.

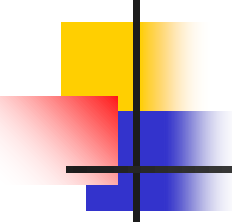


Hierarquia de Classes estruturada como DAG.
Fonte: Carvalho et al, 2007.



Introdução

- Área multidisciplinar
 - Categorização de textos
 - Predição de proteínas
 - Classificação de gêneros musicais e imagens
- O problema da classificação de novos exemplos.

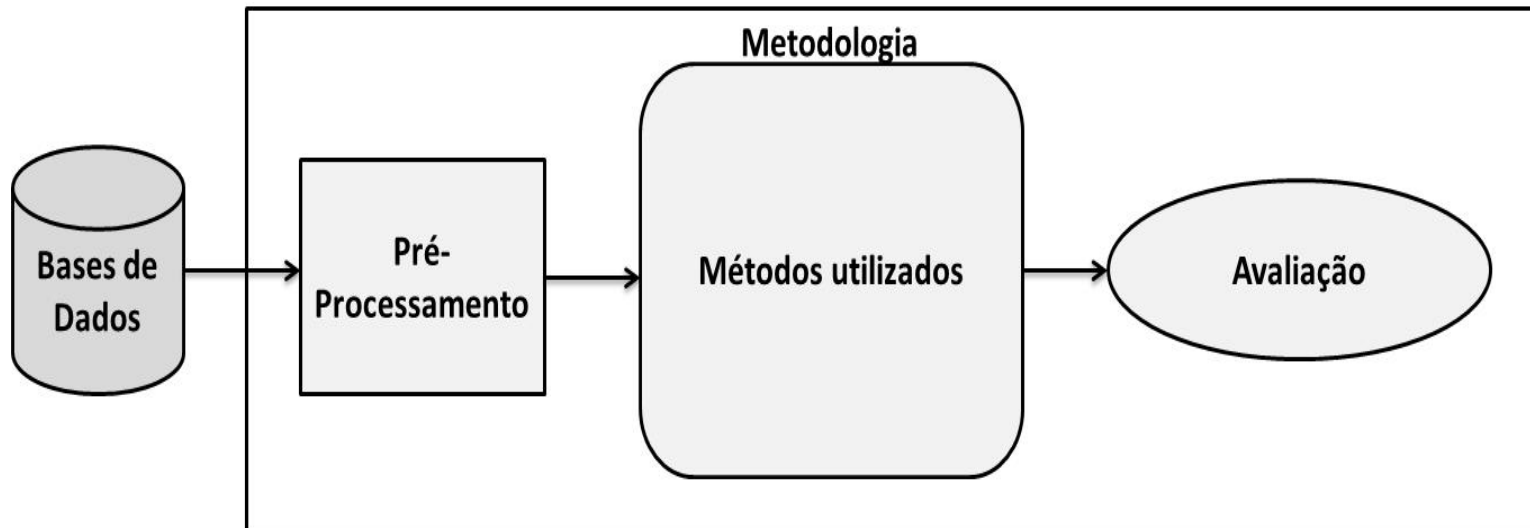


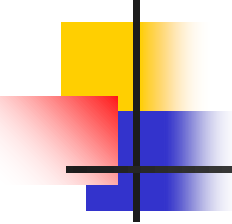
Metodologia

- A metodologia utilizada no trabalho pode ser dividida em duas fases:
 - Aplicação dos métodos em diferentes bases de dados.
 - Avaliação e análise dos resultados.

Metodologia

- Realização dos Experimentos





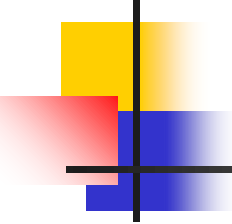
Metodologia

- Base de dados

Tabela 1 - Características das bases de dados GO

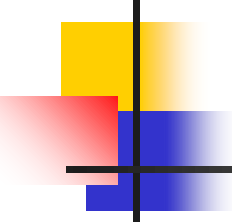
Base de Dados	Quant. Amostras	Quant. Atributos	Quant. Classes	Quant. Max. Níveis	Quant. Min/Max Classes por Amostras	Quant. Min/Max Amostras por Classe
Cellcycle	3751	77	4125	13	3/28	0/785
Church	3749	27	4125	13	3/28	0/786
Derisi	3719	63	4119	13	3/28	0/781
Eisen	2418	79	3573	13	3/28	0/492
Expr	3773	551	4131	13	3/28	0/789
Gasch1	3758	173	4125	13	3/28	0/786
Gasch2	3773	52	4131	13	3/28	0/789
Pheno	1586	69	3127	13	3/28	0/388
Seq	3900	478	4133	13	3/28	0/791
Spo	3697	80	4119	13	3/28	0/775

Fonte: Borges (2012)



Metodologia

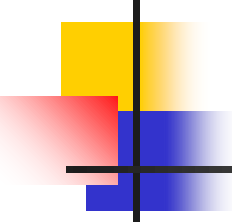
- Para avaliar o desempenho do algoritmo ML-kNN foram implementadas as métricas de precisão e revocação.



Experimentos e Resultados

- Configurações Experimentais
 - As métricas utilizadas foram:
 - Precisão:

$$Prec = \frac{\sum_i |\hat{Y}_i \cap \hat{Y}'_i|}{\sum_i |\hat{Y}_i|}$$



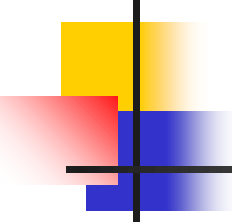
Experimentos e Resultados

■ Configurações Experimentais

■ As métricas utilizadas foram:

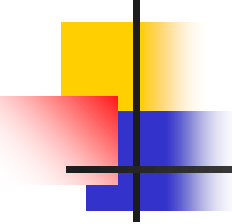
■ Revocação:

$$Rev = \frac{\sum_i |\hat{Y}_i \cap \hat{Y}'_i|}{\sum_i |\hat{Y}'_i|}$$



Experimentos e Resultados

- Configurações Experimentais
 - Variação dos valores de k com 3, 5 e 7
 - Variação dos valores do threshold com 0.5, 0.7 e 0.8
 - Bases divididas conforme a estratégia de holdout



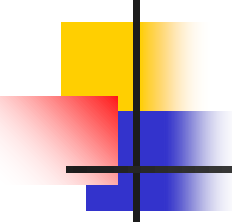
Experimentos e Resultados

■ Resultados – Threshold 0.5

Tabela 2 - Resultados dos experimentos para threshold = 0.5

	Threshold = 0.5					
	K = 3		K = 5		K = 7	
	Prec.	Rev.	Prec.	Rev.	Prec.	Rev.
<u>Cellcycle</u>	0,7359	0,2966	0,7589	0,2888	0,7467	0,2968
Church	0,7687	0,248	0,7721	0,2469	0,7353	0,2631
<u>Derisi</u>	0,7568	0,2573	0,7527	0,2623	0,7606	0,261
<u>Eisen</u>	0,7432	0,2981	0,7406	0,301	0,7445	0,2957
Expr	0,771	0,2926	0,7675	0,2998	0,7597	0,3048
Gasch1	0,7627	0,2915	0,7609	0,2968	0,761	0,2985
Gasch2	0,7474	0,2868	0,7624	0,2778	0,7708	0,274
<u>Pheno</u>	0,7654	0,2441	0,7329	0,2547	0,7431	0,2485
<u>Seq</u>	0,771	0,2864	0,7761	0,2828	0,7797	0,283
<u>Spo</u>	0,741	0,2794	0,7358	0,285	0,7448	0,2802

Fonte: Autoria Própria



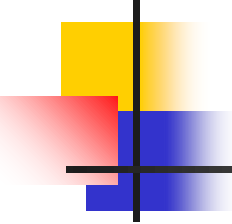
Experimentos e Resultados

■ Resultados – Threshold 0.7

Tabela 3 - Resultados dos experimentos para threshold = 0.7

	Threshold = 0.7					
	K = 3		K = 5		K = 7	
	Prec.	Rev.	Prec.	Rev.	Prec.	Rev.
Cellcycle	0,8986	0,2006	0,8986	0,209	0,9104	0,2013
Church	0,8972	0,1876	0,895	0,1887	0,8973	0,188
Derisi	0,895	0,1908	0,8922	0,1923	0,893	0,1935
Eisen	0,8545	0,2179	0,8707	0,2136	0,8821	0,2075
Expr	0,8841	0,2286	0,8869	0,2282	0,8973	0,2208
Gasch1	0,8809	0,2192	0,8815	0,2287	0,8931	0,2187
Gasch2	0,8985	0,1989	0,8949	0,2071	0,9005	0,2045
Pheno	0,8936	0,1847	0,8922	0,1861	0,8866	0,1865
Seq	0,9121	0,2013	0,9056	0,2062	0,9	0,212
Spo	0,9019	0,1941	0,8897	0,2043	0,8889	0,2005

Fonte: Autoria Própria



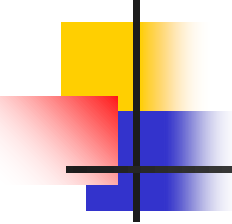
Experimentos e Resultados

■ Resultados – Threshold 0.8

Tabela 4 - Resultado dos experimentos para threshold = 0.8

Threshold = 0.8						
	K = 3		K = 5		K = 7	
	Prec.	Rev.	Prec.	Rev.	Prec.	Rev.
Cellcycle	0,9175	0,1871	0,9245	0,1824	0,9292	0,1834
Church	0,9464	0,1539	0,9467	0,1538	0,9464	0,1538
Derisi	0,9427	0,1563	0,943	0,1554	0,9364	0,1632
Eisen	0,9234	0,1663	0,9056	0,1856	0,9101	0,1857
Expr	0,9243	0,1935	0,9281	0,1948	0,9334	0,1959
Gasch1	0,9201	0,1873	0,9208	0,1951	0,9256	0,1916
Gasch2	0,9195	0,1833	0,9275	0,1811	0,933	0,1792
Pheno	0,9258	0,1604	0,9367	0,1512	0,9296	0,1569
Seq	0,9279	0,1814	0,9179	0,1941	0,9099	0,1934
Spo	0,9398	0,1576	0,926	0,174	0,9276	0,1713

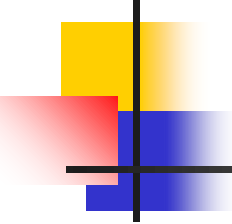
Fonte: Autoria Própria



Experimentos e Resultados

- Análise Estatística

- Para teste de Wilcoxon, a hipótese nula assume que a diferença de desempenho entre algoritmos não é significativa. Com nível de confiança $\alpha=0.05$ a hipótese nula não pode ser rejeitada se $-1.96 \leq z \leq 1.96$.



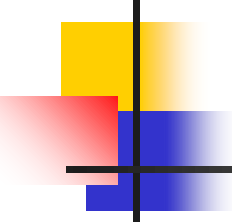
Experimentos e Resultados

- **Análise Estatística**
 - Teste de Wilcoxon com Valor de Corte Constante

Tabela 5 - Teste Wilconxon threshold 0.5

Threshold = 0.5	
K = 3 com k = 5	0.153
K = 3 com k = 7	0.051
K = 5 com k = 7	0.459

Fonte: Autoria Própria



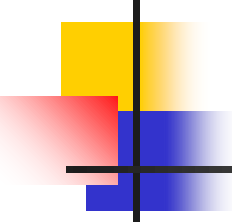
Experimentos e Resultados

- **Análise Estatística**
 - **Teste de Wilcoxon com Valor de Corte Constante**

Tabela 6 - Teste Wilconxon threshold

Threshold = 0.7	
K = 3 com k = 5	0.948
K = 3 com k = 7	0.714
K = 5 com k = 7	1.631

Fonte: Autoria Própria



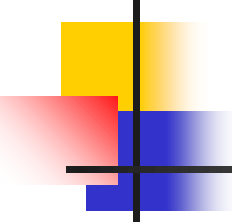
Experimentos e Resultados

- Análise Estatística
 - Teste de Wilcoxon com Valor de Corte Constante

Tabela 7 - Teste Wilcoxon threshold

Threshold = 0.8	
K = 3 com k = 5	0.153
K = 3 com k = 7	0.296
K = 5 com k = 7	0.051

Fonte: Autoria Própria



Experimentos e Resultados

- **Análise Estatística**
 - Teste de Wilcoxon com Valor de k Constante

Tabela 8 - Teste Wilconxon k = 3

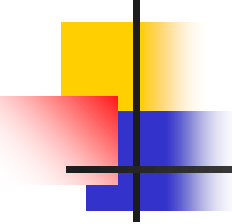
K = 3

Threshold = 0.5 com 0.7 Z-Score = -2.8031

Threshold = 0.5 com 0.8 Z-Score = -2.8031

Threshold = 0.8 com 0.7 Z-Score = -2.8031

Fonte: Autoria Própria



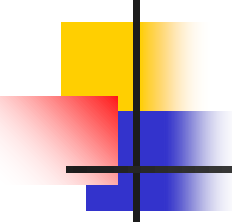
Experimentos e Resultados

- Análise Estatística
 - Teste de Wilcoxon com Valor de k Constante

Tabela 9 - Teste Wilconxon k = 5

K = 5	
Threshold = 0.5 com 0.7	Z-Score = -2.8030
Threshold = 0.5 com 0.8	Z-Score = -2.8030
Threshold = 0.8 com 0.7	Z-Score = -2.8030

Fonte: Autoria Própria



Experimentos e Resultados

- **Análise Estatística**
 - Teste de Wilcoxon com Valor de k Constante

Tabela 10 - Teste Wilconxon k = 7

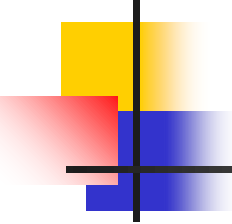
K = 7

Threshold = 0.5 com 0.7 Z-Score = -2.8032

Threshold = 0.5 com 0.8 Z-Score = -2.8032

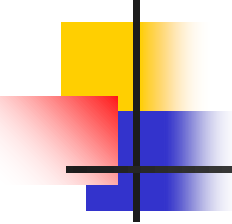
Threshold = 0.8 com 0.7 Z-Score = -2.8032

Fonte: Autoria Própria



Conclusão

- Foram realizadas mudanças para utilizar o algoritmo de classificação multirrótulo ML-kNN.
 - Criação de um novo formato de entrada
 - Aplicação da técnica de Spyromitros
 - Avaliação dos resultados com testes estatísticos



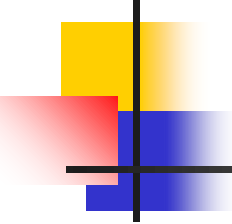
Referências

Dumais, S. and Chen, H. *Hierarchical classification of web content*. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, Greece, pp. 256-263, 2000.

Sun, A. and Lim, E.-P. *Hierarchical text classification and evaluation*. In Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society, pp. 521-528, 2001.

Costa, E. P., Lorena, A. C, Carvalho, A.P.L.F, &Freitas,A. A. (2007). A review of Performance Evaluation Measures for Hierarchical Classifiers. In Proceedings of the AAAI07 – Workshop on Evaluation Methods for Machine Learning II, P.1-6.

Santos, A. M., Canuto, A. M. P. *Investigating the influence of repart in ensemble systems designed by boosting*. In: IJCNN. Hong Kong: IEEE, 2008. P. 2907-2914.



Referências

Spyromitros, E.; Tsoumakas, G.; Vlahavas, I. *An empirical study of lazy multilabel classification algorithms*. In: *Hellenic conference on Artificial Intelligence*, p. 401–406, Berlin, Alemanha, 2009.

Aha, D. W., Kibler, D. e Albert, M. K. (1991). *Instance-based learning algorithms*. *Machine Learning*, 6(1): 37-66.

BORGES, Helyane Bronoski; NIEVOLA, J. C. *Multi-Label Hierarchical Classification using a Competitive Neural Network for Protein Function Prediction*. In: 2012 International Joint Conference on Neural Networks (IJCNN 2012), 2012, Brisbane, Austrália. 2012 International Joint Conference on Neural Networks (IJCNN 2012). Piscataway, NJ: IEEE Press, 2012. v. 1. p. 1-8.