Aprendizado de Maquina para Detecção de *Spam*

Thiago Giroto Milani,
Universidade Estadual Paulista - "Júlio de Mesquita Filho"
Rio Claro – SP
thiago.milani@unesp.br

Fabricio Aparecido Breve,
Universidade Estadual Paulista - "Júlio de Mesquita Filho"
Rio Claro - SP
fabricio.breve@unesp.br

Resumo— Devido a evolução da computação, das técnicas de extração de características (vetorização de texto) e aprendizado de máquina, surgem várias novas aplicações, como na área de segurança da informação para resolver problemas antigos como o recebimento de mensagens de spam. Através da junção dessas duas áreas; aprendizado de máquina (classificação) e a vetorização de texto (extração de características), esse artigo propõe um estudo comparativo entre algoritmos de extração de características e algoritmos de classificação. Alguns resultados preliminares expressivos já foram obtidos com diversos algoritmos.

Área: "Inteligência Computacional".

I. APLICAÇÃO

O crescimento da Internet desde sua criação, com serviços e aplicações cada vez mais abundantes, acabou trazendo também mais oportunidades para que pessoas mal-intencionadas possam disseminar vírus, *malware* e *spam's* com intenção de roubar ou atacar algum usuário ou empresa.

Devido a grande quantidade de mensagens enviadas por *e-mail*, existe a necessidade de se filtrar automaticamente quais são relevantes e quais são propagandas não solicitadas ou tentativas de ataques. Para tanto, fazer a extração das características mais relevantes é uma tarefa importante e um dos pontos principais desta pesquisa, pois estas características são usadas pelos algoritmos de classificação para classificá-las como *spam* ou não *spam*.

Com isso, o objetivo deste projeto de pesquisa é analisar algoritmos de extração de características combinados com classificadores utilizando o aprendizado de máquina, para tentar obter uma melhor identificação das mensagens de *spam*.

II. METODOLOGIA DE DESENVOLVIMENTO

O experimento foi realizado em duas etapas. Inicialmente foi feita a extração de características dos conjuntos de dados sms_collection (SMS), e a spam_ham_dataset (HAM) com os algoritmos TF (*Term Frequency*) e o TFI-DF (*Term Frequency – Inverse Document Frequency*), gerando os vetores com as informações mais relevantes. Posteriormente foi realizada a classificação dos conjuntos com os principais algoritmos de aprendizado supervisionado, e com o algoritmo de Competição e Cooperação entre Partículas (semi-supervisionado) visando identificar qual apresentaria a melhor taxa de acerto e o melhor tempo de execução. Essa fase foi feita usando inicialmente 10% dos itens de dados para treinamento e 90% para testes, e depois com 80% para treinamento e 20% para testes, como mostrado nas tabelas abaixo.

80% DE TREINAMENTO E 20% DE TESTE - ACERTO

ALG,	SPAM	SMS_TF	SMS_IDF	HAM_TF	HAM_IDF
PCC	**97,70%**	98,11%	97,58%	**97,62%**	**97,57%**
LSCV	25,57%	**99,93%**	**99,89%**	84,43%	85,51%
LDA	56,07%	99,39%	98,22%	42,57%	15,44%
LR	76,20%	99,89%	99,82%	86,02%	85,13%
KNN	76,93%	99,95%	99,82%	85,71%	75,01%
CART	72,74%	99,86%	99,86%	84,00%	80,62%
NB	14,37%	99,00%	99,89%	76,36%	86,00%
XGB	81,15%	99,89%	99,89%	86,34%	
RF	77,20%	99,82%	99,82%	85,13%	
ADAB	77,04%	99,87%	99,86%	85,24%	
MLP	76,76%	99,84%	99,82%	87,31%	
SVM	79,13%	99,82%	99,82%	85,13%	

80% DE TREINAMENTO E 20% DE TESTE - SEG,

ALG,	SPAM	SMS_TF	SMS_IDF	HAM_TF	HAM_IDF
PCC	**6**	8	6	**6**	**6**
LSCV	72	**3**	**3**	34	56
LDA	0	441	443	1389	1373
LR	23	8	17	560	3559
KNN	0	346	353	1615	1828
CART	1	2	2	141	223
NB	0	5	6	66	794
XGB	281	275	1311	28171	
RF	0	1	1	7	
ADAB	7	37	37	206	
MLP	9	49	35	480	
SVM	86	12	10	3500	

10% DE TREINAMENTO E 90% DE TESTE - ACERTO

ALG,	SPAM	SMS_TF	SMS_IDF	HAM_TF	HAM_IDF
PCC	**94,91%**	95,32%	95,33%	**95,23%**	**95,05%**
LSCV	21,98%	99,96%	**99,93%**	84,87%	85,53%
LDA	57,30%	98,39%	96,18%	48,92%	19,77%
LR	76,57%	99,95%	99,82%	86,42%	85,13%
KNN	77,30%	99,95%	99,82%	85,86%	75,69%
CART	73,20%	99,96%	99,91%	84,14%	80,95%
NB	12,35%	98,76%	99,93%	75,40%	86,19%
XGB	81,67%	99,95%	99,93%	86,38%	85,13%
RF	77,20%	99,82%	99,82%	85,13%	85,13%
ADAB	77,11%	**99,98%**	99,84%	85,22%	85,13%
MLP	76,87%	99,86%	99,82%	87,23%	85,13%
SVM	79,52%	99,82%	99,82%	85,13%	85,13%

10% DE TREINAMENTO E 90% DE TESTE - SEG.

ALG,	SPAM	SMS_TF	SMS_IDF	HAM_TF	HAM_IDF
PCC	**8**	8	8	**7**	**8**
LSCV	175	6	**7**	75	126
LDA	1	1396	1401	3149	3156
LR	54	16	36	1211	7886
KNN	0	379	389	1778	2015
CART	2	5	5	330	654
NB	0	11	11	111	931
XGB	626	611	3136	63475	1209002
RF	0	1	1	15	14
ADAB	15	**81**	82	443	667
MLP	18	111	75	1011	508
SVM	205	28	22	7929	7476

Os resultados preliminares mostram que o algoritmo de Competição e Cooperação entre Partículas (PCC) com 80% dos itens de dados no treinamento se mostrou mais eficiente em três dos cinco cenários analisados, ficando ainda acima da média nos outros dois, apesar de ser um algoritmo de aprendizado semi-supervisionado, otimizado para trabalhar em cenários com poucos dados rotulados. Com 10% de itens no treinamento, o PCC também se mostra o melhor em três cenários, porém no conjunto de dados *sms_collection* não mostrou tanta eficácia.

Na fase final deste trabalho serão obtidos os resultados que ainda faltam, e também serão incluídos outros algoritmos na comparação, incluindo o OPF (*Optimum-Path Forest*).

REFERÊNCIAS

[1] Cert.br. - Centro de estudos, resposta e tratamento de incidentes de segurança no brasil, cartilha de segurança para internet. Cartilha de Segurança para a Internet, 2016, 2016. URL https://cartilha.cert.br;

[2] M. N. Murty, A. Jain, and P. Flynn. - Data clustering: a review acm compt. surv. ACM Computing Surveys, 31(3), 1999;

[3] V. Moll. - Detecção de intrusão uzando técnica de aprendizado de m aquinas, 2010. Dissertação Mestrado Universidade Federal de Santa Catarina, Florianópolis – SC;

[4] J. de Fernandes Teixeira - Inteligência artificial. Pia Sociedade de São Paulo-Editora Paulus, 2014;

[5] F. A. Breve – Aprendizado de Máquina em Redes Complexas. Tese de Doutorado, Instituto de Cuências Matemáticas e de Computação, Universidade de São Paulo. Citadona, 2010.

IX Workshop do Programa de Pós-Graduação em Ciência da Computação: "Interação entre Academia e Empresa", Unesp, Bauru, 4 e 5 de setembro de 2019.