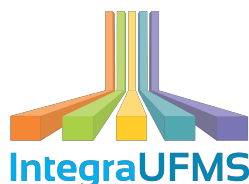




FUNDAÇÃO
UNIVERSIDADE
FEDERAL DE
MATO GROSSO DO SUL



PIBIC UFMS
Programa Institucional de
Bolsas de Iniciação Científica

ESTUDO DE TÉCNICAS ENSEMBLE NO APRENDIZADO BASEADO EM UMA ÚNICA CLASSE NA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTOS

Nathan Dezan¹; Rafael Geraldelli Rossi²;

PIBIC-QSLZW

RESUMO – A classificação automática, pode ser útil para organizar e extrair conhecimento de grandes volumes de textos. Uma forma de viabilizar a classificação automática de textos é por meio de algoritmos de aprendizado de máquina, que possuem a capacidade de aprender, e extrair padrões com coleções de texto [1]. Tradicionalmente é utilizada a classificação multi-classe, na qual um documento pode ser classificado em uma dentre todas as classes, e é necessário apresentar textos rotulados de todas as classes para o algoritmo de aprendizado. Em contrapartida, no aprendizado baseado em uma única classe (ABUC) são fornecidos apenas exemplos da classe de interesse para o algoritmo [2]. Dessa maneira, o algoritmo irá classificar textos como sendo da classe de interesse ou não. Há portanto a diminuição do esforço de rotulação e conhecimento do domínio. Porém, não há um algoritmo capaz obter os melhores resultados em todas situações. Para minimizar este problema, pode-se utilizar técnicas de *ensemble*, as quais combinam os resultados obtidos por diferentes algoritmos, e geralmente produzem resultados iguais ou superiores ao melhor algoritmo individual. Dados os benefícios do ABUC e de *ensembles*, o objetivo deste projeto foi de avaliar o impacto na performance da classificação de textos ao utilizar *ensembles* no ABUC. Mais especificamente, foi utilizada a soma dos votos para aplicar o *ensemble*. Foram obtidos resultados utilizando 4 coleções de textos (CSTR, Classic4, SyskillWebert e Re8), e 4 algoritmos de ABUC (Local Outlier Factor - LOF, One Class SVM - OSCMV, Isolation Forest e Elliptic Envelope), considerando seus resultados individuais e suas combinações no *ensemble*. Os resultados demonstram que a combinação dos algoritmos LOF e OCSVM obtiveram os melhores resultados para a medida F^1 , chegando a superar em até 20% o melhor resultado obtido com um classificador individual.

Palavras-chave: *Ensemble*; Aprendizado Baseado em Uma Única Classe; Classificação de Textos.

¹ Orientador, UFMS.

² Bolsista CNPq (PIBIC): Graduação em Sistemas de Informação, UFMS, nathandezan@ufms.br.

Apoio: CNPq.

Referências

[1] Aggarwal, Charu C. Machine learning for text. Springer, 2018.

[2] Khan, S. S., & Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. The Knowledge Engineering Review, 29(3), 345-374.