Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 5, N. 1, 2017.

Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Comitê de Classificadores *Self Training* para Aprendizado Semissupervisionado

Igor Wescley S. de Freitas[1]
Araken de Medeiros Santos[2]
Matheus da Silva Menezes[3]
Ivan Mezzomo[4]
Departamento de Ciências Exatas, Tecnológicas e Humanas, UFERSA, Campus Angicos, RN

O Aprendizado Semissupervisionado surge baseado no princípio de que não é possível criar um classificador eficaz com poucos exemplos rotulados. Uma pequena parcela dos dados vai nortear a classificação dos exemplos não rotulados, de forma a classificar todo o conjunto [1]. Nessa abordagem, comumente chamada de *Self-Training*(SFT), consiste em tomar como base o conjunto de exemplos rotulados, rotular de forma iterativa e incremental o conjunto de instâncias não rotuladas, e assim gerar um classificador mais efetivo [2].

O escopo do problema esta no fato de que no SFT o erro inicial é propagado até o final do processo, podendo induzir um classificador que atue a partir de exemplos rotulados erroneamente. Outro fato é que o SFT descarta os classificadores intermediários gerados durante o processo. Não existe um mecanismo que considere a atuação contínua dos classificadores intermediários, que podem ter um erro agregado menor.

Os comitês de classificadores destacam-se como estratégia para melhorar o processo de classificação. O desempenho de um classificador é afetado pelos parâmetros usados na indução, como viés indutivo, exemplos usados no treinamento e outros parâmetros do algoritmo, ou seja, parâmetros diferentes produzem classificadores diferentes. Ao invés de ter somente um classificador para rotular, os comitês combinam diferentes classificadores para produzir um resultado mais eficaz.

Na Estratégia original, intuitivamente, o último classificador é o mais preciso, uma vez que foi induzido com um conjunto maior de exemplos rotulados. Todavia, é possível que no início do processo tenha-se produzido muito erro, gerando classificadores não confiáveis. Como proposta de trabalho, foi definida uma estratégia para criar um comitê com arquitetura paralela, formado por todos os classificadores intermediários gerados no *Self-Training*(ESB). Dessa forma, podemos aumentar a eficácia da classificação, combinando todos os classificadores intermediários que possivelmente foram menos afetados pela propagação do erro, minimizando assim os efeitos deste problema.

[1]igorwescley@gmail.com
[2]araken@ufersa.edu.br
[3]matheus@ufersa.edu.br
[4]imezzomo@ufersa.edu.br

Os experimentos foram realizados com as duas estratégias implementadas usando quatro bases: *Splice*, um conjunto formado por pontos de uma sequência de DNA; *SpamBase*, que é uma coleção de e-mails spam; *Satimage*,consiste em valores multi-espectrais de pixels 3×3 vizinhos em imagem de satélite; e *Waveform*, que são registros de ondas fisiológicas e séries temporais. Todas recuperadas do repositório *UCI Machine Learning* [3]. Foram usados também cinco classificadores base: Árvore de decisão, KNN, NaiveBayes, JRip e SVM. Foi definido oito parâmetros de porcentagem de exemplos não rotulados para cada base, que vão de 90% a 20% e e taxa de incorporação por ciclo de 10%. Como módulo combinador foi usado as estratégias de Soma e Voto. O T-test de Student bicaudal [4] com significância de $0,05$, foi utilizado para comparar a relevância estatística entre as estratégias SFT e ESB. Para os testes foi considerado somente o melhor resultado dos cinco classificadores base. Foi considerado também o desempenho por magnitude de cada estratégia.

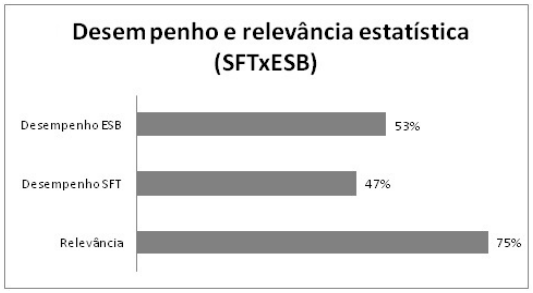

Figura 1: Gráfico com comparação de desempenho e relevância estatística entre SFT e ESB.

A estratégia ESB apresentou melhor desempenho em 53%(17/32) dos casos. Segundo o teste t, 75%(24/32) dos casos são estatisticamente relevantes. O resultado indica que a estratégia tem de fato um melhor desempenho em comparação com a estratégia SFT, e mostra que o uso de um comitê construído com os classificadores intermediários do Self-Training, teve um papel positivo para uma melhor eficácia dos resultados.

Referências

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-trainingt. In *Proceedings of the eleventh annual conference on Computational learning theory (ACM)*, p. 92-100, 1998.

[2] C. Rosenberg, M. Hebert and H. Schneiderman. *Semi-supervised self-training of object detection models*. In 7^o IEEE Workshop on Aplications of Computer Vision, 2005.

[3] UCI Machine Learning Repository at: *http://archive.ics.uci.edu/ml/*. University of California, Irvine, California. Último acesso em 26 de fevereiro de 2015.

[4] G.M. Campos.Estatistica Pratica para Docentes e Pós-graduandos. at: *http://www.forp.usp.br/restauradora/gmc/gmc_livro/gmc_livro_cap19.html*. Último acesso em 14 de março de 2015.