Universidade Federal de Pernambuco
Graduação em Engenharia da Computação
Centro de Informática

2015.2

UM ALGORITMO SUPERVISIONADO PARA GERAÇÃO DE ENSEMBLE

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluna Mariana de Araújo Souza {mas2@cin.ufpe.br}
Orientador George Darmiton da Cunha Cavalcanti {gdcc@cin.ufpe.br}

3 de outubro de 2015

Sumário

1. Contexto 3
2. Objetivo 5
3. Cronograma 6
4. Referências 7
5. Assinaturas 8

1. Contexto

O "No Free Lunch Theorem" [7] leva à conclusão de que não há um algoritmo de aprendizagem de máquina que tenha um desempenho melhor que os outros algoritmos para todos os problemas. Um alternativa para contornar a não existência de um classificador universalmente superior aos outros é combinar vários classificadores, com o intuito de explorar a competência de cada um localmente para obter uma melhor performance no geral [8].

Os Sistemas de Múltiplos Classificadores (MCS) são, então, divididos em três fases: Geração, Seleção e Integração. Na primeira fase, o *pool* de classificadores é gerado. Na segunda fase, um subconjunto não vazio dos classificadores do pool é selecionado. Na terceira e última fase, a resposta dos classificadores selecionados previamente é então combinada para gerar uma resposta final do sistema. Na etapa de Seleção, pode-se ter as seguintes abordagens: Seleção Estática de Classificadores, na qual o mesmo *ensemble* de classificadores é utilizado para classificar todos os padrões de teste, ou Seleção Dinâmica de Classificadores (DES), que por sua vez seleciona ensembles específicos de acordo com cada padrão de teste [2].

Os MCS que utilizam técnicas de DES podem ter altas taxas de acerto apenas com classificadores localmente precisos no pool. Isso pode ser mostrado na Figura 1.1, que ilustra uma base de dados estruturada em uma configuração do tipo XOR. Pode-se observar que, como os dados não são linearmente separáveis, um único Perceptron não pode classificar corretamente as instâncias desse conjunto. Esse exemplo pode ser encontrado em [5].

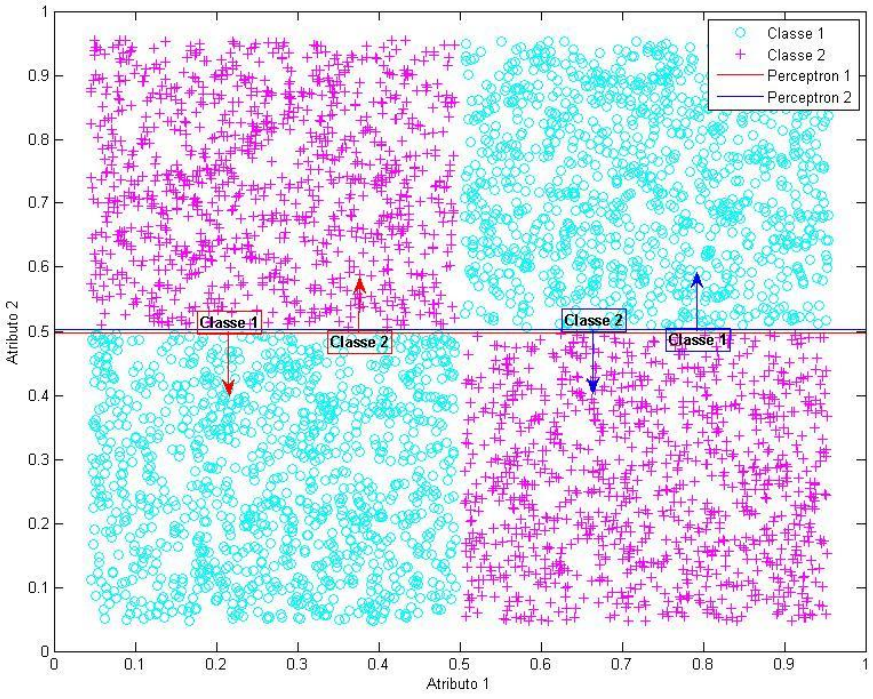

Figura 1.1 - Conjunto de dados representando o problema XOR

Entretanto, esse problema pode ser resolvido utilizando DES. Considerando um pool formado pelo Perceptron 1 e pelo Perceptron 2, cujas respostas são

complementares, cada instância do conjunto de treinamento pode ser corretamente classificada por um classificador no pool. Desse modo, uma taxa de acerto de 100% pode ser obtida, se o Perceptron correto for escolhido na etapa de Seleção.

Este trabalho, portanto, foca na geração de pool de classificadores para MCS que utilizem técnicas de DES. Os métodos clássicos de geração de pool de classificadores são Bagging [1], Random Subspace [3] e Boosting [6], mas muitas variações dos mesmos foram desenvolvidas nos últimos anos.

Bagging é uma técnica na qual um número predeterminado de conjuntos de treinamento são gerados via reamostragem com repetição do conjunto original (método de Bootstrapping). Cada conjunto de treinamento gerado é então usado para treinar um classificador do pool [4,9].

O método de Random Subspace também gera um número predeterminado de conjuntos de treinamento utilizando o conjunto original. Entretanto, cada conjunto formado é um subespaço de menor dimensão do espaço original. Isso é feito reduzindo aleatoriamente o número de atributos do conjunto de treinamento original. Cada classificador do pool é, então, treinado com um dos conjuntos gerados [4,9].

Por outro lado, o método de Boosting e seus variantes são baseados na ideia de que muitos classificadores fracos podem superar em desempenho um classificador forte. A geração desses classificadores fracos é incremental: o algoritmo gera um classificador a cada passo, e cada um deles é treinado com um subconjunto do conjunto de treinamento original. Cada subconjunto é gerado pela atualização de sua distribuição a cada passo, de tal forma que o classificador a ser criado no passo $n+1$ acerte mais as instâncias que o classificador gerado no passo n errou [4,9].

Todos os métodos previamente expostos requerem o treinamento dos classificadores do pool. Mesmo sendo comum utilizar classificadores básicos para reduzir o tempo de treinamento, ainda assim é um processo custoso. Além disso, o número de classificadores no pool tem que ser definido antecipadamente, o que pode não ser vantajoso, já que é razoável acreditar que a quantidade de classificadores necessária para produzir uma boa solução dependa do problema.

2. Objetivo

Este trabalho se propõe a realizar um estudo sobre geração de pool de classificadores em seleção dinâmica de classificadores, com a finalidade de implementar um algoritmo capaz de gerar um pool de Perceptrons. O tamanho desse pool deverá variar de acordo com a distribuição dos dados do conjunto de treinamento, e os Perceptrons gerados não devem ser treinados, com o intuito de otimizar o processo em termos de tempo.

O algoritmo proposto será capaz de produzir um pool tal que cada instância do conjunto de treinamento seja corretamente classificada por pelo menos um dos Perceptrons. Essa abordagem possibilitará que seja gerada a quantidade necessária de classificadores localmente precisos, conforme os dados de treinamento. Além disso, a utilização de uma heurística proposta para obtenção dos pesos dos Perceptrons eliminará a necessidade de treinamento dos mesmos.

3. Cronograma

	Mês															
Atividade	Outubro				Novembro				Dezembro				Janeiro			
Especificação																
Estudo da Bibliografia e Estado da Arte.																
Implementação																
Testes																
Escrita da monografia																
Preparação da apresentação oral																
Apresentação																

4. Referências

[1] Breiman, L. (1996). Bagging predictors. *Mach Learn*, 24(2), pp.123-140.

[2] Britto, A., Sabourin, R. and Oliveira, L. (2014). Dynamic selection of classifiers - A comprehensive review. *Pattern Recognition*, 47(11), pp.3665-3680.

[3] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), pp.832-844.

[4] Kuncheva, L. (2004). *Combining pattern classifiers*. Hoboken, NJ: J. Wiley.

[5] M. O. Cruz, R., Sabourin, R. and D. C. Cavalcanti, G. (2015). *A DEEP analysis of the META-DES framework for dynamic selection of ensemble of classifiers*. [PDF] Cornell University Library, arXiv:1509.00825 [cs.LG].

[6] Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. Boosting the margin: a new explanation for the effectiveness of voting methods, in: *Proceedings of 14th International Conference on Machine Learning*, Nashville, USA, 1997, pp. 322–330.

[7] Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), pp.67-82.

[8] Woźniak, M., Graña, M. and Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16, pp.3-17.

[9] Zhou, Z. (2012). *Ensemble methods*. Boca Raton, FL: Taylor & Francis.

5. Assinaturas

George Darmiton da Cunha Cavalcanti
(Orientador)

Mariana de Araújo Souza
(Aluna)

Recife, 3 de outubro de 2015