Otimização de parâmetros de classificadores para dados de insetos

Bruno Coelho, André Maletzke, Gustavo Batista

ICMC / Universidade de São Paulo
bruno.gomes.coelho@usp.br

Objetivo

Analisar a influência da pesquisa aleatória na otimização de parâmetros dos classificadores, *Support Vector Machines* (SVM) e *Random Forest* (RF) no contexto de identificar insetos.

Métodos e Procedimentos

Utilizando um sensor de luz infravermelha para capturar o voo de insetos, proposto pelo nosso grupo de pesquisa (Silva et al., 2011), adquirimos informações sobre 22 classes de insetos de 16 espécies diferentes, algumas diferenciando entre machos e fêmeas. No total, tem-se 138.323 instâncias, cada uma com 41 atributos. Os dados foram divididos em três conjuntos de dados (CD): Todas as Classes (TC), Mosquitos contra não-mosquitos (M) e *Aedes aegypti* fêmea contra não-fêmea (AF). Para cada um, 70% dos dados foram usados para treinamento e o restante para teste. Utilizando Python e a biblioteca Scikit-learn [1], os classificadores foram treinados com seus parâmetros padrão, gerando SVM-P e RF-P. Para cada tipo de classificador, existem diversos parâmetros que impactam a acurácia. Escolhemos de cada algoritmo dois dos mais importantes para otimizar. Para a SVM Otimizada (SVM-O) a penalidade do termo de erro (C: 0 até 30000) e o coeficiente do *kernel* (CK: 0 até 10). Apenas o *kernel* radial foi utilizado pois uma pesquisa completa sobre seus parâmetros inclui o *kernel* linear (Keerthi et al., 2003). Para a RF Otimizada (RF-O) variamos o número de árvores (A: 10 até 500), e a quantidade de atributos (QA: 1 até o número de atributos) a serem considerados na busca pela melhor divisão. Para variar os parâmetros mencionados acima, realizamos uma pesquisa de grade aleatória com uma validação cruzada de 5-*folds* no conjunto de treino com 60 valores aleatórios (30 para cada parâmetro) para cada classificador. De acordo com Bergstra et al. (2012), isso nos dá 95% de chance de encontrar algo na região do ótimo local. Por último, os melhores parâmetros foram usados para treinar os classificadores sobre todo o conjunto de treinamento.

Resultados

Para os dados TC, M e AF foram encontrados os seguintes valores de parâmetros: para o SVM-O (C:762, KC:1.0789), (C:762, KC:1.0789) e (C:11660, KC:2.7134), e para o RF-O (NT:215, NF:35), (NT:227, NF:18) e (NT:376, NF:28).

Tabela 1: Acurácia para cada classificador e CD.

	SVM		RF	
	D	O	D	O
TC	93.82%	**97.07%**	95.87%	**97.31%**
M	97,82%	**99.01%**	98.94%	**99.07%**
AF	88.52%	**91.57%**	92.70%	**95.99%**

Conclusões

Para ambos os classificadores observou-se que a pesquisa aleatória fornece resultados melhores que a configuração utilizando parâmetros padrão.

Referências

Wu, X., et al. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Silva, D.F., et al. (2011). Resultados Preliminares na Classificação de Insetos Utilizando Sensores Óticos.

Keerthi, S.S., & Lin, C.J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. Neural computation, 15(7), 1667-1689.

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.