26º SIICUSP

Optimizing classifier parameters for insect datasets

Bruno Coelho, André Maletzke, Gustavo Batista

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

Objective

Analyze the influence of a random search for optimizing the parameters of two of the most recommended machine learning classifiers, Support Vector Machines (SVM) and Random Forest (RF) in the context of identifying insects.

Materials and Methods

Utilizing an infrared light sensor designed by our research group for flying insects (Silva et al.), we acquired data on 22 classes of insects from 16 species, some considering male or female as separate classes. In total, we have 138,323 instances, each one with 41 features. The data was then divided into 3 independent problems, All Classes (AC), Mosquitoes versus non-mosquitoes (M) and *Aedes aegypti* female versus non-female (AF). For each one, 70% of the data was user for training while the remaining 30% was used for testing. Using Python and the Scikit-learn library [1], classifiers were trained with the default parameters, generating SVM-D and RF-D. For each classifier algorithm, there are many parameters that impact the accuracy. We choose two of the most important for each algorithm for tuning. For the SVM Optimized (SVM-O) the penalty of the error term (C: 0 to 30000) and the kernel coefficient (KC: 0 to 10). The Radial kernel was used since a complete search over the parameters encompasses the linear Kernel (Keerthi et al., 2003). For the RF Optimized (RF-O) we varied the number of trees (NT: 10 to 500), and features (NF: 1 to the number of features) to consider when looking for the best split. To vary the parameters mentioned above a random grid search using 5-fold cross-validation of the training data was used to check 60 random values (30 for each parameter) for each classifier. According to Bergstra et al. (2012) this gives a 95% chance of finding something in the local optimum. The same initial seed was used across all the problems. Finally, the best parameters estimated were used for training on the whole training set and evaluated on the test set.

Results

The SVM-O and RF-O had the following parameters for the datasets AC, M, and AF, respectively: (C:762, KC:1.0789), (C:762, KC:1.0789) and (C:11660, KC:2.7134), while the RF-O had (NT:215, NF:35), (NT:227, NF:18), and (NT:376, NF:28).

Table 1: Accuracies for each classifier and dataset.

	SVM		RF	
	D	O	D	O
AC	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%**

Conclusions

For both classifiers, the random search can be seen to give better results than the default parameters.

References

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.