DOI 10.1016/j.compag.2017.01.007

Ensembles of wrappers for automated feature selection in fish age classification

Sergio Bermejo

Departament d'Enginyeria Electrònica,

Universitat Politècnica de Catalunya (UPC),

Jordi Girona 1-3 (C4 building), 08034 BARCELONA, SPAIN

PHONE: +34 4016758, E-MAIL: sergio.bermejo@upc.edu

Abstract. In feature selection, the most important features must be chosen so as to decrease the number thereof while retaining their discriminatory information. Within this context, a novel feature selection method based on an ensemble of wrappers is proposed and applied for automatically select features in fish age classification. The effectiveness of this procedure using an Atlantic cod database has been tested for different powerful statistical learning classifiers. The subsets based on few features selected, e.g. otolith weight and fish weight, are particularly noticeable given current biological findings and practices in fishery research and the classification results obtained with them outperforms those of previous studies in which a manual feature selection was performed.

Keywords: Automated fish age classification, Atlantic cod otoliths, feature selection, nearest neighbor classifiers, statistical pattern recognition, support vector machines.

1. Introduction

One of the most challenging problems in the field of pattern recognition (PR) is feature extraction (Guyon et al., 2006), which aims finding the most compact and discriminative set of properties or “features” presented in data. Although many research in feature extraction has been addressed to automate such a process, it has traditionally been considered a task much more problem- or domain-dependent than others in PR (Duda et al., 2001) since a good knowledge of the domain could be used to obtain such features, at least tentatively.

Fish age classification, a PR task of vital relevance among others for stock assessment and management (Girdler et al., 2010), usually relies on such manual procedures for feature extraction. In this direction, several fish features have been proposed for use in statistical fish age prediction and classification, with special emphasis in recent years to fish otolith features based on Fourier descriptors (Fablet and Le Josse, 2005; Galley et al., 2006) and different morphological parameters (Burke et al., 2008; Bermejo et al., 2007; Robotham et al., 2010; Hua et al., 2012).

However, the generalization error of statistical classifiers –i.e. their ability to mistake new examples taken on the same problem– tends to increase as of the number of features (Raudys and Jain, 1991) and, accordingly, the use of an arbitrary number of them leads to poor performance. One example of such behavior was demonstrated in (Bermejo, 2014) using multi-class support vector machines for fish age classification of an Atlantic cod database. Hence, if automatic feature extraction methods were additionally employed for reducing the complexity of the feature space a better performance could presumably be obtained. Other important benefits of such strategy includes speeding up computation (e.g. decreasing training times) and data understanding or reverse engineering (i.e. to increase knowledge about the problem, which can be of vital significance in natural sciences like fisheries science).

While some authors (e.g. Webb, 2002) consider feature extraction a process only concerning transformation of the original variables, it is generally agreed that feature extraction comprises the following steps: feature construction or generation that performs some kind of preprocessing –e.g. a linear or non-linear transformation– of the original raw variables (Theodoridis and Koutroumbas, 2008) and feature selection (Guyon and Elisseeff, 2003) that chooses a subset of the original or transformed variables.

There are three main approaches to feature selection (Blum and Langley, 1997; Guyon and Elisseeff, 2003, 2006): filter methods, wrappers and embedded methods. While filters can be viewed as a preprocessing step since they select a subset of variables independently of the chosen predictor (e.g. a classifier), wrappers use it as a black box or subroutine to score subsets of variables and embedded methods perform variable selection in its training phase. In this way, wrappers are based on an arguably better estimate of accuracy obtained with the predictor that will employ the feature subset than a separate measure that may have a completely unrelated inductive bias, but, at the expense of a higher computational cost (Blum and Langley, 1997). However, the inherent variance (or instability) of feature subset selection methods (Guyon and Elisseeff, 2006) produces a plethora of very different subsets attained for different conditions, i.e. different parameter tuning, small perturbations of the dataset or presence of redundant features.

In this paper, a novel wrapper that use a form of ensemble learning (Dietterich, 2003), which are based on a strategic combination of several predictors, have been proposed to attain a greater stabilization and thus a better generalization of the feature selection process. Feature subsets obtained with the ensemble of wrappers which employ as base classifiers support vector machines and nearest neighbor classifiers allow achieving a classification performance that outperforms a previous study

(Bermejo, 2014). Moreover, these subsets that have very few features, e.g. only otolith weight and fish weight, are of relevance in accordance with recent findings in fisheries research.

2. Materials and methods

2.1. Atlantic cod database

This dataset contains morphological and biological features for codfish age classification. Traditional methods for determining the age of fish usually focus on analyzing hard parts of the body, such as otoliths, which are small particles in the inner ear composed of a gelatinous matrix and calcium carbonate, since the macroscopic growth patterns of otholiths are correlated with the fish' age.

The fish database consists of one hundred forty-five Atlantic cod of known age (varying from two to six years) from the Plateau stock that were hatched the same year and later kept and reared in pen cages. This dataset was created from originally fish of known-age sampled at different years in captivity since a number of samples were recaptured once a year. Otoliths were taken from this stock and weighed and also four morphological features were recorded following an image analysis method defined in (Bermejo et al., 2007). Additionally, fish length, weight and sex were available for each sample.

The leave-one-out (LOO) error using a 1-NN rule (Devroye et al., 1996; pp. 407-421) were computed for this set (19.31%) as a way to estimate the Bayes error, i.e. the minimum amount of classification error achievable. In a previous study with this database using SVMs (Bermejo, 2014), the minimum obtained error was 21.79% for otolith weight, fish length, weight and sex acting as features, which is lower than an error rate of 22% obtained for a related dataset, combining five experts' readings, who were given low and intermediate levels of information about fishes and the conditions that they were obtained (Doering-Arjes et al., 2008). According to the above considerations, some improvement in accuracy is still possible with SVMs taking the value of the LOO estimate as an approximate lower bound to the attainable misclassification rate. Table 1 displays the results of the LOO estimate and also includes other relevant information of this dataset. A more comprehensive description of the cod database is presented in (Bermejo, 2014).

2.2. Ensemble of wrappers

Ensemble learning methods, such as bagging, boosting and variants (Bauer and Kohavi, 1999) are based on the formation of a set of predictors $\{\varphi(\boldsymbol{x};\boldsymbol{D}_k)\}$ trained on a sequence of learning sets {$\boldsymbol{D}_k$}, which are typically generated from a single dataset $\boldsymbol{D}$ using a resampling technique such as bootstrapping (Efron and Tibshirani, 1994). The second core element of any ensemble method is a combination strategy: the most obvious and effective procedure for combining a sequence of K predictors $\{\varphi_k\}$ whose outputs are continuous is averaging (Breiman, 1996a), i.e. $\bar{\varphi}=\sum_k \varphi_k / K$. Ensembles have been built specifically to select features; for example, variants of AdaBoost for feature selection have been proposed using decision stumps (Long and Vega, 2003) and a mutual information measure (Liu et al., 2008), random subspace methods have also been employed in feature ranking for removal of irrelevant variables (e.g. Tuv et al., 2009), and ensembles based on bootstrapping have been combined with recursive feature elimination and feature ranking (Windeatt et al., 2007). Furthermore, several studies have analyzed the use of averaging and voting for the combination of multiple feature selection criteria with the hope that several criteria would reflect different properties in feature subsets (e.g. Somol et al., 2009), although none of them has analyzed the effect of these procedures using a sole criterion to obtain a single feature subset. Our proposal addresses this problem in the context of wrappers.

Wrappers (Kohavi, 1995) select features from a pool of feature sets based on a decision rule of the form $\varphi_W = \arg\min_j L_{CV}\left(\mathrm{C}_{D}^{j}; \boldsymbol{D}\right)$, that is, they select the jth feature set for which $L_{CV}\left(C_{D}^{j}; \boldsymbol{D}\right)$ is the minimum, where L_{CV} is the cross-validation error based on the dataset ***D*** computed in the base classifier $C_{D}^{j} = C\left(\boldsymbol{x}^{j}; \boldsymbol{D}\right)$, whose inputs belong to the jth feature set space. If the database is divided into a learning set ***D*** for performing cross-validation and a test set ***T*** for final assessment of the classifier after feature selection, a sequence of learning sets {$\boldsymbol{D}_k$} and test sets {$\boldsymbol{T}_k$} can be generated for different random splits of the database. Then, and in accordance to the theoretical analysis given in (Breiman, 1996a, 1996b), we propose in this paper a stabilized feature selection rule that can be obtained through averaging over L_{CV} in order to stabilize the metric used in wrappers directly, so the feature selection rule based on an ensemble of wrappers (EW) can be computed as $\overline{\varphi}_{EW} = \arg\min_j \left(\sum_k L_{CV}\left(C_{\mathbf{D}_k}^{j}; \mathbf{D}_k\right)\big/K\right)$. The proposed stabilization of the assessment criterion can be simply seen as an averaging of several *k*-fold cross-validation estimates (based on the output of the wrapper's base classifier) similarly to the way in which the outputs of several classifiers are stabilized through averaging. The reader is referred to Breiman, 1996a, 1996b for further discussion, and definition, of stability.

A baseline algorithm for feature selection with wrappers using internal cross-validation (Flach, 2012) is suggested in Algorithm no. 1. The ensemble approach using rule $\overline{\varphi}_{EW}$ is detailed in Algorithm no. 2 as a straightforward variation of the baseline algorithm, in which feature selection is postponed until all the splits obtained in the first version are evaluated. In this way, the second algorithm uses the same amount of computational resources than the first one but a single decision on what features are more relevant is obtained averaging over all these splits.

2.3. Base classifiers

Reducing the instability of the base classifiers would make it possible to evaluate the degree of stability achieved by $\overline{\varphi}_{EW}$ with respect to φ_W and could also provide additional insight into how the stabilized decision rules work. Specifically, if the induction algorithm $C^j_{\mathrm{D}_k}$ is completely stable on a sequence of learning sets {$\boldsymbol{D}_k$}, then $C^j = C(\boldsymbol{x}^j; \boldsymbol{D}_i) = C(\boldsymbol{x}^j; \boldsymbol{D}_k)\ \mathit{for}\ \forall i,k$. Thus, the metric $\sum_k L_{CV}(C^j; \boldsymbol{D}_K)/K = \overline{L}_{CV}(C^j)$, where $\overline{L}_{CV}$ denotes an averaged form of the cross-validation error computed using different random replicates of the original database. As K augments, $\overline{L}_{CV}$ will use more samples from the database than L_{CV}, which is based on a single replicate, and can thus presumably obtain a better estimation. Following this rationale, two well-known stable induction algorithms, SVMs and NNs, have been employed as base classifiers in wrappers.

SVMs (Vapnik, 1998), which has been developed in accordance with main results of statistical learning theory, have also obtained a practical success in a range of practical problems that makes them an appreciated part of many practitioners' toolbox. Multi-class SVMs (Hsu and Lin, 2002) are a required extension of two-class SVMs that deal with R-class classification problems, with R>2. In the experiments, we used two multi-class SVMs implemented in the Spider library (Weston et al., 2006): 1) 1-vs-R ("one-against-all") SVMs (Steinwart and Christmann, 2008), and 2) 1-vs-1 ("one-against-one") SVMs (Schölkopf and Smola, 2001). Other SVM algorithms also implemented in the library were ruled out in a previous round of experiments, since the results obtained with them were outperformed by both 1-vs-R and 1-vs-1 SVMs.

Nearest-neighbor classifiers (Duda et al., 2001; pp.161-214) remain one of the simplest yet most valuable nonparametric classification procedures. Given a set of labeled prototypes ***P***, the *k*-NN algorithm assigns the test point ***x*** to that class majority among its *k* nearest neighbors belonging to ***P*.** In the experiments reported, the 1-NN, also simply denoted as the NN rule, was used, since it has less computational burden than the *k*-NN rule. Although the NN rule is sub-optimal with respect to the *k*-NN rule in terms of the asymptotic error probability (i.e. with an unlimited number of prototypes), its error rate is never worse than twice the Bayes error (Devroye et al., 1996; pp. 61-90).

2.4. Statistical assessment of experiments

As pre-processing, whitening –i.e. mean removal and scaling by the variance of each feature– was performed on the dataset so as to prevent the negative effect of their very different scaling on the SVMs and NNs, and thus improving dramatically their classification accuracy (see e.g. Ali and Smith-Miles, 2006). In (Bermejo, 2014), the positive effect of such standardization is specifically discussed for this dataset.

A previous round of simple experiments was done to limit the set of values for the parameters of the multi-class SVMs. According to the results obtained, radial basis function (RBF) kernels $K(\mathbf{x},\mathbf{x}_i)=\exp\left(\|\mathbf{x}-\mathbf{x}_i\|^2 / 2\sigma\right)$ were selected with a kernel width of $\sigma=\{5,10,15,20,25\}$, while the rest of the parameters involved were the default values defined in the Spider library (Weston et al., 2006).

The whole training set was chosen as nearest-neighbor prototypes in order to reduce the computational burden due to the use of the learning algorithm. This brute-force strategy, which usually works better than significant condensing and editing, achieves competitive results with learning algorithms that compute a reduced number of prototypes (see e.g. Bermejo, 2000).

Since the datasets here are medium- and small-sized, it was considered preferable to maximize the learning set size in order to get enough training data. Thus, test sets were formed containing only 25% of the database the test set size according to common practices found in the literature; in particular, test sets ranged from 50% to 25% of the complete database in fourteen datasets from the STATLOG project (Michie et al., 1994). Accordingly, the datasets were first randomly divided using stratification into a test set $\boldsymbol{T}_i$ (25%) and a learning set $\boldsymbol{D}_i$ (75%) for each split i=1,…,I of the database (with I=75 when SVMs are used as the base classifiers and *K*=100 for NNs). Then, $\boldsymbol{D}_i$ was divided using stratification into five equal-sized parts or folds (i.e. n=5) that maintained approximately the original proportion of data belonging to each class; in order to reduce variance in the estimates of classification accuracy, this random division of $\boldsymbol{D}_i$ was repeated ten times, forming a sequence of folds. Thus, steps 5-13 of Algorithms 1 and 2 were repeated ten times and results conveniently averaged; in the case of SVMs, a sequence of classifiers using a kernel width of σ={5,10,15,20,25} was also generated for each split i, each feature set j and fold, and only those classifiers with parameters obtaining, on average, the best results on the validation set were retained for testing. Finally, the relative frequency with which the rule $\overline{\varphi}_{EW}$ outperforms or equals φ_W defined by $\gamma = \sum_i 1\left(Err_i\left(\overline{\varphi}_{EW};\mathbf{T}_i\right) \leq Err_i\left(\varphi_W;\mathbf{T}_i\right)\right)\big/ I$ was computed in order to compare Algorithms 1 and 2.

3. Results and discussion

As Table 2 shows, on average, the use of $\overline{\varphi}_{EW}$ improves accuracy, since $Err\left(\overline{\varphi}_{EW}\right) < Err\left(\varphi_W\right)$ for all the classifiers (see also Fig. 1). Also, for each data split i, feature selection done by averaging mainly improves the results achieved by classifiers based on feature sets selected using cross-validation, since $\gamma \in [.75,.96]$ (see also Fig. 2).

While the feature selection rule $\overline{\varphi}_{EW}$ generates a single feature set (see Table 2), φ_W generates a population of feature sets, which only sometimes coincides with $\overline{\varphi}_{EW}$ (these cases are shown as points along the line depicted in Fig. 2). On the other hand, feature sets obtained by $\overline{\varphi}_{EW}$ are not unique with respect to the problem, but depend on the wrapper's base classifier. However, although there is not a total consensus among the classifiers, features set obtained by the selection rule $\overline{\varphi}_{EW}$ are particularly coherent with biological findings, since fish weight (W) and otolith weight (OW) –i.e. the features selected when 1-vs-R SVMs are used as base classifiers– and fish length (L), which is also included when NN classifiers are used, are known to be highly correlated with age and are often used in automatic fish age estimation or classification (Lou et al., 2005, 2007; Metin and Ilkyak, 2008; Ochwada et al., 2008; Pino et al., 2004), although other researchers have proposed the use of other features, such as otolith growth rings (Fablet and Le Josse, 2005; Guillaud et al., 1999, 2000; Rodin et al., 1996) or otolith shape (Bird et al., 1986; Campana and Casselman, 1993; Castonguay et al., 1991). Additionally, and more importantly, the feature set obtained by the selection rule $\overline{\varphi}_{EW}$ (based only on OW and W) in combination with 1-vs-R SVMs achieves an average test error (20,93%) that outperforms best results computed with previous SVM experiments (Bermejo, 2014) with the same dataset in which feature set selection was performed manually (21,79%).

The feature selection rule $\overline{\varphi}_{EW}$ makes it possible to compute a single feature set with the additional information obtained by generating different splits of the original database. Since the repetition of experiments for different splits seems to be recommended to reduce variance in test results (at least for small databases), $\overline{\varphi}_{EW}$ can be used in this context at no extra computational cost. In order to extend this procedure to datasets with a greater number of features, the brute-force search can be replaced with the inspection of a pool of candidates obtained by ordering the feature set space by leave-one-out error, since the minimum leave-one-out errors are obtained for feature sets quite similar to those computed by $\overline{\varphi}_{EW}$ (see Table 1). Also, search strategies (Guyon, 2006; pp.119-136) applied to large dimensionality domains in the context of wrappers (Gheyas and Smith, 2010) are useful for obtaining a feature set subspace where $\overline{\varphi}_{EW}$ and the experimental procedure suggested above were run with moderate computational resources.

4. Conclusions

A metric based on averaging, a well-known method employed in ensemble learning for stabilizing, has been proposed to reduce the instability of the feature subset selection process performed by wrappers and has been tested on an Atlantic cod dataset using SVMs and NN classifiers as base classifiers. As shown, a single feature subset can be obtained in such a form of ensemble of wrappers and used to reverse engineer or better explain data. Features selected in fish age classification are particularly noticeable in view of current biological findings and practices in fishery research and outperforms SVM classification accuracies obtained with manual feature selection (Bermejo, 2014).

Acknowledgements

The author would like to thank the Institute of Marine Sciences of Barcelona (ICM-CSIC) for supplying the otolith images used in this work. Also, it is acknowledged the funding received in the EU's IBACS project under contract no. QQLRT-2001-01610.

References

Ali, S., Smith-Miles, K.A., 2006. Improved support vector machine generalization using normalized input space. In: Sattar, A., Kang, B.H. (Eds.), Advances in Artificial Intelligence Lecture Notes in Computer Science, vol. 4304. Springer-Verlag, Berlin, pp. 362–371.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36, 105-139.

Bermejo, S., 2000. Learning with Nearest Neighbour Classifiers. Ph.D. Dissertation, Departament d'Enginyeria Electrònica, Universitat Politècnica de Catalunya.

Bermejo, S., Monegal, B., Cabestany, J., 2007. Fish age categorization from otolith images using multi-class support vector machines. Fish. Res. 84, 247-253.

Bermejo, S., 2014. The benefits of using otolith weight in statistical fish age classification: A case study of Atlantic cod species. Comp. Elec. Agr. 107, 1–7.

Bird, J.L., Eppler, D.T., Checkley, D.M., 1986. Comparisons of hearing otoliths using Fourier series shape analysis. Can. J. Fish. Aquat. Sci. 43, 1228–1234.

Blum, A. L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271.

Breiman, L., 1996a. Heuristics of instability and stabilization in model selection. Ann. Statist. 24, 2350-2383.

Breiman, L., 1996b. Bagging Predictors. Mach. Learn. 24, 123-140.

Burke, N., Brophy, D., King, P.A., 2008. Shape analysis of otolith annuli in Atlantic herring (*Clupea harengus*); a new method for tracking fish populations. Fish. Res. 91, 133–143.

Campana, S.E., Casselman, J.M., 1993. Stock discrimination using otolith shape analysis. Can. J. Fish. Aquat. Sci. 50, 1062–1083.

Castonguay, M., Simard, P., Gagnon, P., 1991. Usefulness of Fourier analysis of otolith shape for Atlantic mackerel (*Scomber scombrus*) stock discrimination. Can. J. Fish. Aquat. Sci. 48, 296–302.

Devroye, L., Györfi, L., Lugosi, G., 1996. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, Berlin.

Dietterich, T. G., 2003. Ensemble learning. In: M. A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, second edition. The MIT Press, Cambridge, MA, pp. 405–408.

Doering-Arjes, P., Cardinale, M., Mosegaard, H., 2008. Estimating population age structure using otolith morphometrics: a test with known-age Atlantic cod (*Gadus morhua*) individuals. Can. J. Fish. Aquat. Sci. 65, 2342–2350.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification (2nd Edition). Wiley-Inter-science, New York.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman & Hall, London.

Fablet, R., Le Josse, N., 2005. Automated fish age estimation from otolith images using statistical learning. Fish. Res. 72, 279 –290.

Flach, P. A., 2012. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, Cambridge.

Galley, E. A., Wright, P. J., and Gibb, F. M., 2006. Combined methods of otolith shape analysis improve identification of spawning areas of Atlantic cod. ICES J. Mar. Sci. 63, 1710-1717.

Girdler, A., Wellby, I., Welcomme, R., 2010. Fisheries Management: A Manual for Still-Water Coarse Fisheries. Wiley-Blackwell, Oxford.

Gheyas, I. A., Smith, L. S., 2010. Feature subset selection in large dimensionality domains, Pattern Recognit. 43, 5-13.

Guillaud, A., Ballet, P., Troadec, H., Rodin, V., Benzinou, A., Le Bihan, J., 1999. A multiagent system for edge detection: an application to growth ring detection on fish otoliths, in: Image Processing and its Applications, 1999 Seventh International Conference Publication No. 465, Vol. 1, pp. 445-449.

Guillaud, A., Troadec, H., Benzinou, A., Rodin, V., Le Bihan, J., 2000. Continuity perception using a multiagent system: an application to growth ring detection on fish otoliths, In: Pattern Recognition, 2000. Proceedings. 15th International Conference, Vol. 2, pp. 519-522.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157-1182.

Guyon, I., Elisseeff, A., 2006. An Introduction to Feature Extraction. In: I.Guyon, M. Nikravesh, S. Gunn, L. A. Zadeh (Eds.), Feature Extraction Foundations and Applications. Studies in Fuzziness and Soft Computing, Vol. 207, Springer, New York, 2006, pp.1-26.

Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L. A., (Eds.), 2006. Feature Extraction Foundations and Applications. Vol. 2007, 1-778. Springer, New York.

Hua, J., Li. D., Duan, Q., Han, Y., Chen, G., Si, X., 2012. Fish species classification by color, texture and multi-class support vector machine using computer vision. Comput. Electron. Agr. 88, 133–140.

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multi-class support vector machines. IEEE Trans Neural Netw. 13, 415-25.

Kohavi, R., 1995. Wrappers for performance enhancement and oblivious decision graphs. Ph.D Thesis, Department of Computer Science, Stanford University.

Liu, T.-Y., Li, G.-Z., Yang, J.Y., Yang, M.Q., 2008. Feature selection for the imbalanced QSAR problems by using EasyEnsemble. Int. J. Comp. Biol. Drug Des. 1, 334-346.

Long, P. M., Vega, V. B., 2003. Boosting and microarray data, Mach. Learn. 52, 31–44.

Lou, D.C., Mapstone, B.D., Russ, G.R., Davies, C.R., Begg, G.A., 2005. Using otolith weight–age relationships to predict age-based metrics of coral reef fish populations at different spatial scales. Fish. Res. 71, 279–294.

Lou, D.C., Mapstone, B.D., Russ, G.R., Davies, C.R., Begg, G.A., Davies, C.R., 2007. Using otolith weight–age relationships to predict age based metrics of coral reef fish populations across different temporal studies. Fish. Res. 83, 216–227.

Metin, G., Ilkyak, G.M., 2008. Use of otolith length and weight in age determination of poor cod (*Trisopterus minutus* Linn., 1758). Turk J. Zool. 32, 293–297.

Michie, D., Spiegelhalter, D.J., Taylor, C.C., 1994. Machine Learning, Neural and Statistical Classification, Prentice Hall, New York. Available online at: http://www.amsta.leeds.ac.uk/~charles/statlog/.

Ochwada, F.A., Scandol, J.P., Gray, C.A., 2008. Predicting the age of fish using general and generalized linear models of biometric data: a case study of two estuarine finfish from New South Wales, Australia. Fish. Res. 90, 187–197.

Pino, C.A., Cubillos, L.A., Araya, M., Sepúlveda, A., 2004. Otolith weight as an estimator of age in the Patagonian grenadier, *Macruronus magellanicus*, in central-south Chile. Fish. Res. 66, 145–156.

Raudys, S. J., Jain, A. K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE T. Pattern. Anal. 13, 252-264.

Robotham, H., Bosch, P., Gutiérrez-Estrada, J. C., Castillo, J. & Pulido-Calvo, I., 2010. Acoustic identification of small pelagic fish species in Chile using support vector machines and neural networks. Fish. Res. 102, 115–122.

Rodin, V., Troadec, H., de Pontual, H., Benzinou, A., Tisseau, J., Le Bihan, J., 1996. Growth ring detection on fish otoliths by a graph construction, in: Proceedings of the International Conference on Image Processing, Vol. 1, pp. 685-688.

Schölkopf, B., Smola, A. J., 2001. Learning with Kernels. The MIT Press, Cambridge.

Steinwart, I., Christmann, A., 2008. Support Vector Machines. Springer, Berlin.

Somol, P., Grim, J., Pudil, P., 2009. Criteria Ensembles in Feature Selection, in: J.A. Benediktsson, J. Kittler, F. Roli (Eds.): MCS 2009, LNCS 5519, Spinger, Berlin, pp. 304–313.

Theodoridis, S., Koutroumbas, K., 2008. Pattern Recognition, Fourth Edition. Academic Press, New York.

Tuv, E., Borisov, A., Runger, G., Torkkola, K. (2009). Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination, J. Mach. Learn. Res. 10, 1341-1366.

Vapnik, V., 1998. Statistical Learning Theory, Wiley Inter-science, New York.

Webb, A. R., 2002. Statistical Pattern Recognition, 2nd Edition. Wiley, New York.

Weston, J., Elisseeff, A., BakIr, G., Sinz, F., 2006. The Spider (matlab toolbox). Available online at: http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html.

Windeatt, T., Prior, M., Effron, N., Intrator, N., 2007. Ensemble-based Feature Selection Criteria, in: Petra Perner (Ed.): Machine Learning and Data Mining in Pattern Recognition (MLDM 2007), IBaI publishing, Leipzig, pp. 168-182.

Algorithm 1 Baseline algorithm for wrappers based on internal cross-validation

1: For i=1,…,I

2: Split database randomly into a test set $\mathbf{T}_i$ and a learning set $\mathbf{D}_i$ using a ratio 1:q where 1:q denotes the sampling ratio between $\boldsymbol{D}_k$ and $\boldsymbol{T}_k$, i.e. % of samples q/(1+q) is sampled for $\boldsymbol{D}_k$ and % 1/(1+q) for $\boldsymbol{T}_k$

3: For j=1 to 2^m combinations of feature sets

4: Obtain for feature space j^{th} a subset $\mathbf{D}_i^j$ from $\mathbf{D}_i$ where j is a vector in a binary representation $(j_1 \cdots j_m)$ with j_k denoting whether feature k^{th} is present ('1') or not ('0') and $\mathbf{D}_i^j \in X^p, \mathbf{D}_i \in X^m, \mathbf{D}_i^j \subset \mathbf{D}_i, \mathbf{D}_i^j \in \Re^p, \mathbf{D}_i \in \Re^m, 0 < p \le m$

5: Split $\mathbf{D}_i^j$ into n disjoint sets $\{\mathbf{D}_i^{j,k}, k=1,...,n\}$, i.e. $\bigcup_{k=1}^{n} \mathbf{D}_i^{j,k} = \mathbf{D}_i^j, \bigcap_{k=1}^{n} \mathbf{D}_i^{j,k} = 0$

6: For k=1 to n folds

7: Obtain a training dataset $\mathbf{D}_i^{j,-k} = \bigcup_{m=1,\, m\neq k}^{n} \mathbf{D}_i^{j,m}$ and a validation set $\mathbf{V}_i^{j,k} = \mathbf{D}_i^{j,k}$

8: Define a sequence of classifiers' parameters $\{\boldsymbol{\sigma}_l, l=1,..., L\}$

9: For l=1,…,L

10: Compute classifier $C_l(\mathbf{x}^j; \mathbf{D}_i^{j,-k}, \boldsymbol{\sigma}_l)$ or, in short, $C_l(\mathbf{x}^j; \boldsymbol{\sigma}_l)$, i.e. a classifier $C_l(\mathbf{x}^j)$ working in feature space X^p with $\mathbf{x}^j \in X^p$ using the training data set $\mathbf{D}_i^{j,-k}$ for the classifier's parameters $\boldsymbol{\sigma}_l$

11: Obtain the cross-validation error for $C_l(\mathbf{x}^j; \boldsymbol{\sigma}_l)$ as the loss error for this classifier computed using $\mathbf{V}_i^{j,k}$, i.e. $L_{CV}(C_l^j) = L(C_l(\mathbf{x}^j; \boldsymbol{\sigma}_l), \mathbf{V}_i^{j,k})$

12: Choose the best classifier $C^k(\mathbf{x}^j)$ of the sequence $\{C_l\}$ with optimal parameters $\boldsymbol{\sigma}^k$ as the one that minimizes the cross validation (CV) error, i.e.
$C^k(\mathbf{x}^j; \boldsymbol{\sigma}^k) = \arg_C \min_l L_{CV}(C_l^j)$ or $L_{CV}(C^{k,j}) = \min_{l=1,...,L} L_{CV}(C_l^j)$

13: Obtain mean CV error in $\mathbf{D}_i^j$ for feature space j^{th} as $L_{CV}(\mathbf{D}_i^j) = \frac{1}{n}\sum_{k=1}^{n} L_{CV}(C^{k,j})$

14: Select the feature subset from which the mean CV error $L_{CV}(\mathbf{D}_i^j)$ is minimum, i.e. $\varphi_W(i) = \arg\min_j L_{CV}(\mathbf{D}_i^j)$

15: Obtain the generation error $Err_i(\varphi_W(i); \mathbf{T}_i)$ of classifiers in feature space $\varphi_W(i)$

16: Compute the mean generalization error for the baseline wrapper φ_W as

$$Err(\varphi_W) = \sum_{i=1}^{I} Err_i(\varphi_W(i); \mathbf{T}_i) \Big/ I$$

Algorithm 2 Ensembles of wrappers (as a variation of Algorithm 1)

1: For i=1,…,I
2: Split database randomly into a test set $\mathbf{T}_i$ and a learning set $\mathbf{D}_i$ using a ratio 1:q
3: For j=1 to 2^m combinations of feature sets
4: Obtain for feature space jth a subset $\mathbf{D}_i^j$ from $\mathbf{D}_i$ with

$$\mathbf{D}_i^j \in X^p, \mathbf{D}_i^j \in X^m, \mathbf{D}_i^j \subset \mathbf{D}_i, \mathbf{D}_i^j \in \Re^p, \mathbf{D}_i \in \Re^m, 0 < p \leq m$$

5: Split $\mathbf{D}_i^j$ into n disjoint sets $\{\mathbf{D}_i^{j,k}, k = 1,..., n\}$
6: For k=1 to n folds
7: Obtain $\mathbf{D}_i^{j,-k} = \bigcup_{m=1,m\neq k}^{n} \mathbf{D}_i^{j,m}$ and $\mathbf{V}_i^{j,k} = \mathbf{D}_i^{j,k}$
8: Define a sequence of classifiers' parameters $\{\boldsymbol{\sigma}_l, l = 1,..., L\}$
9: For l=1,…,L
10: Compute classifier $C_l\left(\mathbf{x}^j; \mathbf{D}_i^{j,-k}, \boldsymbol{\sigma}_l\right)$
11: Obtain $L_{CV}\left(C_l^j\right) = L\left(C_l\left(\mathbf{x}^j; \boldsymbol{\sigma}_l\right), \mathbf{V}_i^{j,k}\right)$
12: Choose $C^k\left(\mathbf{x}^j; \boldsymbol{\sigma}^k\right) = \arg_C \min_l L_{CV}\left(C_l^j\right)$ or

$$L_{CV}\left(C^{k,j}\right) = \min_{l=1,...,L} L_{CV}\left(C_l^j\right)$$

13: Compute $L_{CV}\left(\mathbf{D}_i^j\right) = \frac{1}{n}\sum_{k=1}^{n} L_{CV}\left(C^{k,j}\right)$
14: For i=1,…,I
15: Compute the mean CV error for feature space jth as $L_{CV}(j) = \frac{1}{I}\sum_{i=1}^{I} L_{CV}\left(\mathbf{D}_i^j\right)$
16: Select the feature subset from which the mean cross-validation $L_{CV}(j)$ is minimum, i.e. $\varphi_{EW} = \arg\min_j L_{CV}(j)$
16: For i=1,…,I
17: Obtain the generation error of classifiers in feature space φ_{EW} for $\mathbf{T}_i$ as $Err_i\left(\varphi_{EW}; \mathbf{T}_i\right)$
18: Compute the mean generalization error for the averaged wrapper φ_{EW} as

$$Err\left(\varphi_{EW}\right) = \sum_{i=1}^{I} Err_i\left(\varphi_{EW}; \mathbf{T}_i\right) \Big/ I$$

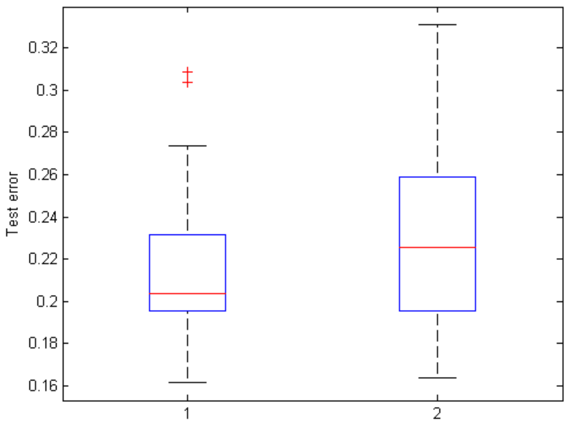
Size	No. of Features	Features / Feature vector	No. of Classes	Minimum Leave-one-out Error
145	8	Fish sex (S), fish length (L), fish weigh (W), otolith weight (OW), otolith contour length (C), otolith area (A), otolith maximum internal distance (I), otolith maximum perpendicular distance (P) / (P I A C OW W L S)	5 [fish age: 2 to 6]	0.1931 [for feature set $12=(00001100)_2$]

Table 1. Codfish dataset summary.

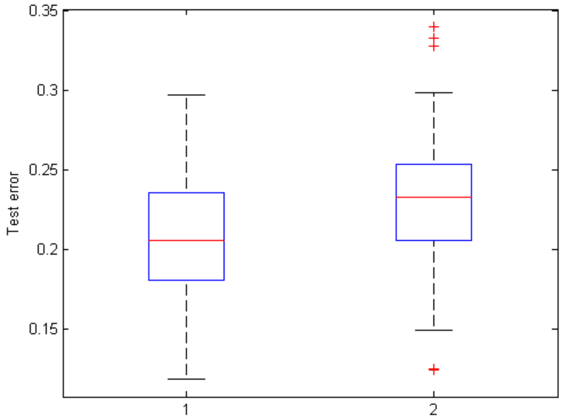
		$Err(\varphi_W)$	$Err(\overline{\varphi}_{EW})$	Feature vector(*) / $\overline{\varphi}_{EW}$	γ
SVM	1-vs-1	.2297	.2147	(P I A C OW W L S)/ 175=$(10101111)_2$	.74567
	1-vs-R	.2273	.2093	(P I A C OW W L S)/ 12=$(00001100)_2$	.96
NN		.2459	.214	(P I A C OW W L S)/ 14=$(00001110)_2$	.84

Table 2. Comparison of feature set selection using averaging and cross-validation.

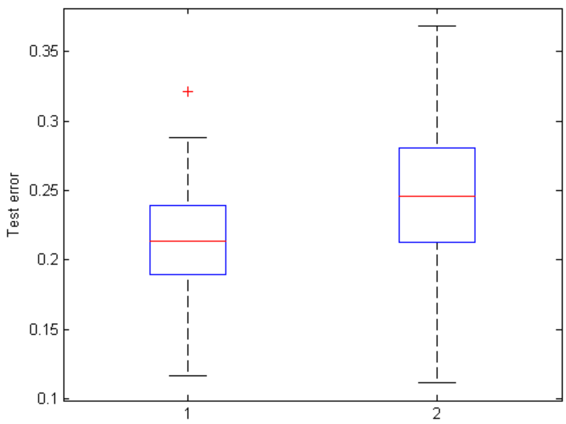
(* see Table 1 for further details)



a)

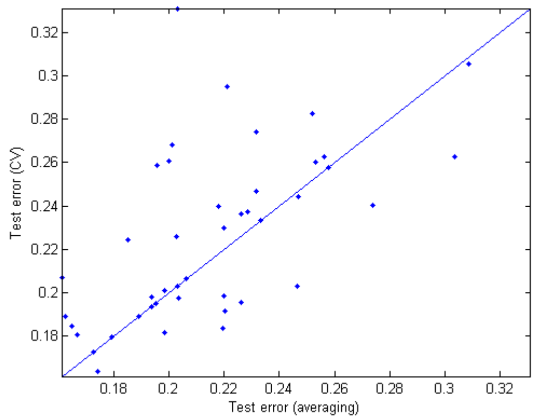


b)

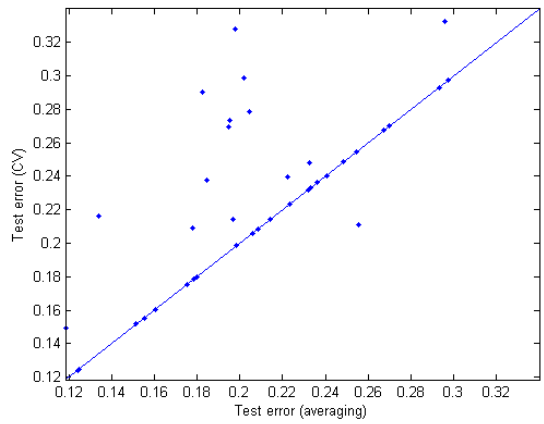


c)

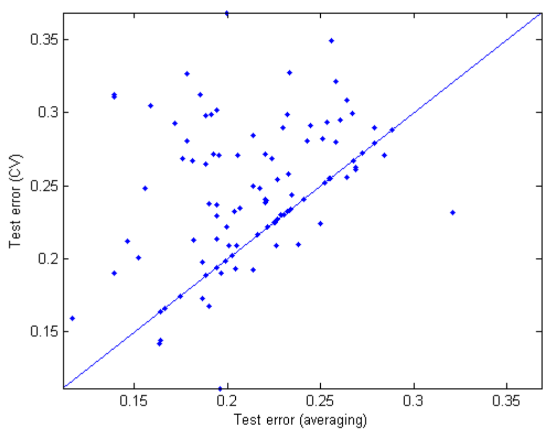
Fig.1. Box plot of average test errors $Err(\bar{\varphi}_{EW})$ [left] and $Err(\varphi_W)$ [right] using: a) 1-vs-1 SVMs, b) 1-vs-R SVMs and c) NN classifiers.



a)



b)



c)

Fig.2. Test errors of ensembles of wrappers based on averaging, $Err_i(\overline{\varphi}_{EW})$, vs. those based on internal CV, $Err_i(\varphi_W)$, for different $\mathbf{T}_i$ using a) 1-vs-1 SVMs, b) 1-vs-R SVMs and c) NN classifiers.