Journal of Physics: Conference Series

An ensemble of dissimilarity based classifiers for Mackerel gender determination

To cite this article: A Blanco *et al* 2014 *J. Phys.: Conf. Ser.* **490** 012130

An ensemble of dissimilarity based classifiers for Mackerel gender determination

A. Blanco, R.Rodriguez, I. Martinez-Maranon

AZTI-Tecnalia Astondo Bidea, Edificio 609 - Parque Tecnológico de Bizkaia - 48160 Derio (Bizkaia)

E-mail: ablanco@azti.es

Abstract.
Mackerel is an infravalored fish captured by European fishing vessels. A manner to add value to this specie can be achieved by trying to classify it attending to its sex. Colour measurements were performed on Mackerel females and males (fresh and defrozen) extracted gonads to obtain differences between sexes. Several linear and non linear classifiers such as Support Vector Machines (SVM), k Nearest Neighbors (k-NN) or Diagonal Linear Discriminant Analysis (DLDA) can been applied to this problem. However, theyare usually based on Euclidean distances that fail to reflect accurately the sample proximities. Classifiers based on non-Euclidean dissimilarities misclassify a different set of patterns. We combine different kind of dissimilarity based classifiers. The diversity is induced considering a set of complementary dissimilarities for each model. The experimental results suggest that our algorithm helps to improve classifiers based on a single dissimilarity.

1. Introduction

Mackerel is a pelagic specie plentiful in the Northeast atlantic waters. The idea of distinguishing and selecting fish species is justified by the high value of roes compared with the price of the fish. The method chosen is based on the color difference between the gonads of male and female fish consists in a sensor that allows to characterize the spectrum obtained after the incidence of electromangetic radiation into the gonads

A variety of machine learning techniques have been proposed such as SVM, k NN or DLDA. However they rely on the use of the Euclidean distance that fails often to reflect accurately the proximities among samples. The classifiers have been extended to work with non-Euclidean dissimilarities. In spite of this, the resulting algorithms misclassify a different set of patterns. Combining non-optimal classifiers can help to reduce particularly the variance of the predictor. To achieve this goal, different versions of the classifier are usually built by sampling the patterns or the features. Nevertheless, this kind of resampling techniques reduce the size of the training set. We build the diversity of classifiers considering three models and a set of complementary dissimilarities for each model. Finally, the classifiers are aggregated using a voting strategy.

2. Combination of dissimilarity based classifiers

An important step in the design of a classifier is the choice of a proper dissimilarity that reflects the proximities among the objects.

SVM and DLDA are not able to work directly from a dissimilarity matrix. The SVM algorithm is extended to work from a dissimilarity matrix by defining a kernel of dissimilarities. DLDA is adapted following a different approach by embedding the patterns in a Euclidean space.

The DLDA is a variant of the Linear Discriminant Analysis with diagonal and constant covariance matrices. A vectorial representation of the data should be obtained.

Our method builds the diversity of classifiers considering three different kind of models such as SVM, k-NN and DLDA and several dissimilarities.

Our combination algorithm proceeds as follows: A set of complementary dissimilarities are computed. For the SVM algorithm, the kernel of dissimilarities is computed and the optimization problem is solved in the usual way. k-NN is able to work directly from a dissimilarity matrix. DLDA algorithm, the dissimilarities should be embedded in an Euclidean space via a Multidimensional Scaling algorithm. The ensemble of classifiers is aggregated by a standard voting strategy.

3. Results and Discussion

Mackerel in their best maturity stages from Basque fishing ports was used to the test the ensemble. Colour measurements was performed on females and males extracted gonads. The reflectance spectra were recollected and twelve colour parameters were analyzed to characterize differences between both sexes.

Two datasets were considered. The first consisted on fresh mackerel samples. 1006 samples (from 2007-2010). The second dataset consisted on 1439 samples of defrozen Mackerel.

The dissimilarities have been computed without normalizing not to increase the correlation among them.The algorithm chosen to train the SVM is C-SVM. The C regularization parameter has been set up by ten fold-crossvalidation. We have considered non-linear kernels. The number of neighbors for k-NN algorithm is estimated by cross-validation.

Table 1. Empirical results for the best single classifier for each technique.

Technique	Datasets	Error %	False negative %	False positive %
SVM RBF (Correlation)	Fresh	7.05%	5.16%	1.88%
SVM RBF (Manhattan)	Defrozen	2.51%	1.47%	1.04%
K-NN (Manhattan)	Fresh	8.84%	5.16%	3.67%
K-NN(Chi-squared)	Defrozen	2.51%	0.86%	1.65%
DLDA (Correlation)	Fresh	8.74%	2.98%	5.76%
DLDA (Correlation)	Defrozen	3.67%	0.19%	3.47%
	Fresh	6.16%	3.280%	2.88%
Combination	Defrozen	1.78%	0.19%	1.58%

The algorithms have been evaluated considering the global errors and the false negative and positive errors. Both have been estimated by ten-fold cross-validation

Table 1 shows the experimental results for the best single classifier for each technique and the proposed technique and the following conclusions can be drawn:

- The dissimilarity that minimizes the error depends strongly on the classifier and on the particular dataset considered. No dissimilarity outperforms the others for a wide range of models and datasets. Hence the choice of a proper dissimilarity is not an easy task for human experts.
- The combination strategy proposed outperforms the misclassification errors of the best single classifiers. In particular, the ensemble of classifiers improves significantly the SVM

algorithms for the two datasets considered. We also report that our method improves the best k-NN classifier for Fresh and Defrozen Mackerel. Finally, DLDA is also improved for both datasets.

4. Conclusions

A new classification scheme that consider different dissimilarities and different classifiers have been proposed. The algorithm has been applied to the Mackerel gender determination in fresh and defrozen species Our method helps to reduce the misclassification errors of the best single classifier for both datasets. As future research trends, a new classification scheme will be developed to take into account different classes of kernels a nd new combination strategies will be applied.

[1] Blanco A, Manuel Martn-Merino M , De Las Rivas J 2008, *Innnovations in Hybrid Intelligent Systems*. 393 – 400.
[2] Dudoit S,Fridlyand J, and Speed T 2002, *Journal of the American Statistical Association*, B **97** 77–87.
[3] Furey T, Cristianini N, Duffy N, Bednarski D, Schummer M, and Haussler D 2000 *Bioinformatics*. B **16** 906–914.
[4] Kuncheva L I 2004. Pattern Classifiers.
[5] Lucio P 1997 *Biology and Behaviour*. B **21** 1–23 .
[6] Pekalska E , Paclick P and Duin R 2001, of Machine Learning Research. B **2** 175–211, .
[7] Valentini G and Dietterich T 2004 *Journal of Machine Learning Research*. B **5** 725–775.
[8] Vapnik V 1998 *Statistical Learning Theory* .