A Machine Learning Approach to Predict the Pink Shrimp Harvest in the Patos Lagoon Estuary

A Machine Learning Approach to Predict the Pink Shrimp Harvest in the Patos Lagoon Estuary

Paulo Drews-Jr.[1], Matheus Bauer[1], Karina Machado[1], Pedro Puciarelli[2] and Luiz Felipe Cestari Dumont[2]

[1] Centro de Ciências Computacionais, Universidade Federal do Rio Grande - FURG, Brazil
[2] Instituto de Oceanografia, Universidade Federal do Rio Grande - FURG, Brazil

Abstract. This paper presents a novel methodology to predict the natural behavior of pink shrimp (Farfantepenaeus paulensis) harvest, in the Patos Lagoon Estuary (PLE) by using supervised machine learning. This prediction is a critical task due to its environmental, economic and social impact. Supervised machine learning algorithms such as Support Vector Machines (SVM), decision trees and rules learning were combined with meta-learning techniques to perform the discrete prediction of the harvest. Performance of several classifiers is evaluated by a set of metrics, especially by a specific metric to deal with the inherent relation of order between the classes. The official harvest data, provided by government agencies, may be affected by random and systemic errors caused mainly by illegal fishing and lack of efficient landing control. These errors, together with the lack of knowledge of the fishing effort employed, increase the difficulty of the prediction task. Results obtained using meta-learning techniques combined with classic algorithms reached an accuracy of 91% for the pink shrimp harvest prediction.

Categories and Subject Descriptors: J.3 [**Computer Applications**]: Life and Medical Sciences; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Shrimp Prediction, Meta Learning, Supervised Learning

1. INTRODUCTION

Pink shrimp - *Farfantepenaeus paulensis* - is an important fishery resource in the South and Southeast regions of Brazil and the main fishery resource captured by artisan fishing in the Patos Lagoon Estuary (PLE) [D'Incao 1991], located on Rio Grande do Sul, the southernmost State of Brazil. High variability of the harvest size is a known characteristic of *penaeidae* [Diop et al. 2007], family in which the pink shrimp is included. According to [Paiva 1997], Rio Grande do Sul is the fourth Brazilian State in terms of landings originated from artisan fishing, also being the largest producer of pink shrimp in this country, accounting for more than 40% of the Brazilian production. The PLE's production represents more than 90% of the artisan fishery captured in Rio Grande do Sul [D'Incao and Reis 2002]. Therefore, the knowledge on the harvest behavior is essential either for fishermen and government. The application of supervised machine learning techniques and algorithms to analyze environmental variables, related to the pink shrimp harvest, is an alternative to the increase of knowledge of this important process in the PLE.

Supervised machine learning techniques are divided according to how the input data is analyzed and classified. Techniques such as decision trees, lazy learning, Support Vector Machines (SVM) and rules learning are applied to a database and evaluated to find which one has the best prediction capacity over the analyzed data. An alternative to the use of a single technique at time is to use a set of techniques, known as meta-learning [Kotsiantis 2007]. These techniques learn using meta-data

This research is partly supported by CNPq, CAPES and FAPERGS. This paper also represents a contribution of the INCT-Mar COI funded by CNPq Grant Number 610012/2011-8. We also would like to thank the colleagues from LCD/IO-FURG for helping with the experimental data.

provided by supervised classifiers to obtain the classification.

The prediction of biological data is a complex task due to several factors related to the origin, manipulation and availability of these data. Data acquired at field can be incorrectly acquired due to inappropriate equipment operation or even by small accuracy. On the machine learning context, an additional difficulty is the lack of balance between the classes to be predicted, *i.e.* a class having much more examples than others interfere on the learning capacity of the algorithms.

Based on a time series data provided by several sources, the main objective of this work is to predict the behavior of the pink shrimp harvest in the PLE. The main contributions are the methodology to predict the behavior of the ELP and the evaluation of a set of classifier in this problem.

2. RELATED WORK

It is possible to find on a literature review of this field, researches that predict population abundance by using biological and/or environmental data as input for models, that range from simple equations to advanced machine learning approaches.

For instance, a modeling based on Markov Chains is proposed by [Grant et al. 1988] to predict the annual harvest of shrimp in the Gulf of Mexico. Their database was built through simulation and composed by attributes related to volume of capture, fishing area, depths, natural and fishing mortality. Good predictions were possible on June and July. However, the complexity of the phenomena and data availability are different from the present work.

Using a eighteen year time-series (1986-2004) from the Charleston Harbor (South Carolina, USA) [Garcia et al. 2007] predicted the white shrimp harvest as catch-per-unit-effort (CPUE) and state commercial fisheries landing (SCFL). Feed-forward artificial neural networks were used with SCFL, CPUE, salinity and temperature as input data for one (t+1) and three (t+3) months ahead predictions. The methods proposed by [Garcia and Almeida 2006] defined the employed delay of the attributes. [Garcia et al. 2007] reached 92% of accuracy for SCFL(t+1) and 79% for SCFL(t+3) according to the Spearman's correlation coefficient [Spearman 1987]. The prediction of the pink shrimp harvest in the PLE requires more attributes than the used by [Garcia et al. 2007], due to the more complex recruitment pattern of these species, involving oceanographical, meteorological and biological phenomena.

[Sujjaviriyasup and Pitiruek 2013] compared machine learning techniques: ARIMA [Box et al. 2008], Holt-Winters model [Kalekar 2004] and SVM [Vapnik 2006]. The analyzed databases were white shrimp harvest and chicken production. Both data from Thailand from January/2007 to December/2012. The databases were split in training (70%) and test (30%) subsets, and Support Vector Machines (SVM) was found to be the most accurate technique. Results using this technique is presented in the present work.

A vital information to achieve successful management in a fishery is to know the amount of stock production. The pink shrimpâĂŹs fishery has a direct impact on the local economy, therefore, it is important to predict how it may occur on the following year. To the best of our knowledge, there is no harvest prediction methodology, of any species for the Patos Lagoon Estuary (PLE), thus the use of supervised machine learning techniques is an innovative approach on the context of the estuary's study. The pink shrimp harvest presents direct impacts over the local economy and society.

3. METHODOLOGY

In this section we covered all aspects related to our proposed methodology to predict the behavior of the pink shrimp harvest in the PLE. The tools, techniques and algorithms are described, as well the origin and the data selection.

3.1 Source of the Data

Several organizations contributed with the data we used to build our predict models. Firstly, the attribute to be predicted (harvest of pink shrimp) is provided by the Brazilian Environmental Agency (IBAMA) on kilograms by month. The harvest is the most critical variable due to the large delay to make available this data.

The IO-FURG provided data from the Long Term Ecological Research (PELD) project, which began in 1996. The biological data were monthly obtained, from eight sampling sites. Information, such as water temperature, salinity, and the number of shrimps caught, by manual towing net, were used. All the data provided by the PELD project represents an average of five from the eight collection points. These points are selected according to the frequency of occurrence of pink shrimp.

Wind is appointed by [D'Incao 1991] as a key attribute to the abundance of pink shrimp. The wind requires two component to be correctly represented, *i.e.* magnitude and direction. The Brazilian Institute of Meteorology (INMET) provided monthly wind direction (categorized using INMET Code) and wind speed (m/s).

National Oceanic and Atmospheric Administration (NOAA) is a North American agency specialized on climatic data, for this work the agency provided the *El Niño* anomaly index associated to Region 3. Table I shows all the attributes used on the models with their minimum and maximum values.

Table I. Attributes used by our methodology and their range of valid values.

Attribute	Minimum Value	Maximum Value
Water salinity (PSU)	0	31.75
Water temperature (^0C)	8.0	31.94
Mean of captured shrimps ($\mathbb{R}$)	0.0	15.75
Mean of captured shrimps (length <= 80mm) ($\mathbb{R}$)	0.0	501.92
El Niño anomaly on region 3 (NOAA Anomaly index)	-1.81	3.62
Wind Direction (INMET Code)	0	36
Wind Speed (m/s)	1.3	5.84

The seven attributes showed on Table I are the result of a selection from, initially, twenty attributes. Multiple factors were used to perform the attribute selection task, among them are the high reliability, the availability and release frequency of the data. Moreover, some attributes were removed due to not cover all the analyzed period. Attributes which were no longer being updated and/or released were replaced by the corresponding attribute from another source.

3.2 Time Series Analysis

The PELD project began on August/1996 with monthly expeditions, which still occurs nowadays. The data from the pink shrimp harvest is organized in months by the IBAMA, with the last release on December/2011. This release limits our availability of data.

The fishing season of the pink shrimp is determined according to investigations on several biological aspects, but mainly the size of the shrimps in the estuary [D'Incao 1991]. The fishing is allowed from the beginning of February to late May. In fact, analyzing the harvest data, by month, from August/1996 until December/2011 these four months represented more than 98% of the total harvest: February (13.5%), March (38%), April (30.7%) and May (16.2%). According to [D'Incao 1991], the shrimp, on a early stage of life, enters in the estuary by the end of September with more intensity on October and November.

Due to the bigger capture of the pink shrimp on the period from February to May, these four months were chosen as the target months, *i.e.* the months which the harvest is predicted. Furthermore, these months is allowed to harvest on the ELP, while the other months are prohibited. The pink shrimp presents a complex cycle of life that is related with the environmental and ecological conditions from months before the harvest. For each month to be predicted (February to May), data from the previous four months are used on the analysis. For example, data from November, December, January and February are used to predict a March harvest. Considering four harvests per year, a time series of fifteen years is composed by sixty harvests, from the end of 1996 to the begin of 2011.

3.3 Harvest Categorization

The idea to convert from numerical harvest data to categorical prediction is due the limitation of reliability of the provided data. The official harvest data, provided by government agencies, can be affected by random and systemic errors caused mainly by illegal fishing and lack of efficient landing control. Thus, the numeric value does not provided much more information than the categorical in this context. The knowledge of the perspective of harvest in the PLE is very important due to the previously mentioned reasons, thus the categorical prediction is capable to provide enough information for the governmental agency and fisherman to control the fishery efforts and insurance issues.

Initially, the numerical attribute the harvest are categorized on three distinct categories (bad, regular or good). This task is based on either specialists expertise and statistics estimation.

A bad harvest is defined as a harvest with less than 1 tonne (1000kg/month). Considering the standard deviation of the non-null harvests of all months from August/1996 to December/2011 the values for regular and good harvests is defined. From 1 tonne to 638 tonne the harvest is defined as regular, above that as a good harvest. After apply the the categorization on the sixty harvests, we obtained twelve bad harvest, forty regular and eight good harvests.

3.4 Classifiers

As previously described, the use of meta-learning techniques is motivated by the fact that lower performance algorithms could provide additional information regarding the analyzed subject [Kittler et al. 1998]. Therefore, the use of an ensemble of classifiers could compensate some individual weakness of each individual algorithm.

3.4.1 *Conventional Supervised Learning Algorithms.* The SVM algorithm is described by [Vapnik 2006] on which the simplest case (linear) a hyperplane is built to separate positive and negative examples with a maximum margin. The use of kernel functions enables to deal with non-linear data. We adopted the polynomial kernel because provides the best results.

Algorithms based on rules learning have on their simple outputs an advantage, mainly when the specialist the knowledge about relevant attributes is required. The JRip classifier is an implementation of the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm, proposed by [Cohen 1995], where rules are created through tests with all the possible values for all available attributes, seeking gain of information.

The DecisionTable is another algorithm of the rule learning category, described by [Kohavi 1995a]. This technique maps the classes through rules built using values or range of values of the attributes. For a non classified instance, the DecisionTable seeks a rule that fits perfectly to the instance, returning the correspondent class, if a perfectly combination is not found, the most frequently class of the table is then returned.

The REPTree classifier creates regression trees based on information gain and variance reduction, which are reached using the pruning method for the error reduction and sorting the numeric attributes once [Göndör and Breffelean 2012].

Another decision tree algorithm is the DecisionStump, described by [Holte 1993], which creates a tree with only one node. The tree is generated by the algorithm interruption when the most informative attribute is added to the tree. According to [Iba and Langley 1992] the DecisionStump algorithm can reveal interesting characteristics of a dataset, in spite of their simplicity and the number of suppositions made.

The Locally Weighted Learning (LWL) is a lazy learning and memory based learning technique. Lazy learning techniques store the training data and only do training when a new instance needs to be classified [Atkeson et al. 1996], a distance function is used to find on the training data a similar distance to the non classified one [Frank et al. 2003].

3.4.2 *Meta-Learning Algorithms.* According to [Vilalta and Drissi 2002], meta-learning algorithms investigates how the learning systems have their performance increased by the experience. The repeated application of a conventional learning algorithm over the same database produces always the same result, although with meta-learning algorithms the results may change.

The Voting technique combines the predictions from a set of classifiers, named base classifiers, through a given rule [Kuncheva 2004]. The rule can be the median, minimum or maximum probability, majority voting, product or average of probabilities, among others.

The Grading algorithm [Seewald and Fuernkranz 2001] train base classifiers, for each base classifier one meta classifier is trained whose task is to predict whether and when the base classifier will err. The original database receives another column representing a new class, which inform if the base classifier did a correctly or incorrectly prediction.

Described by [Breiman 1996], the Bagging algorithm trains multiples classifiers in order to make predictions. A training set is generated for each experiment and the final classifier is formed by the set of classifiers generated on the experiments. An instance is then classified through the vote of all trained classifiers for each class.

3.5 Performance Evaluation

Besides of the classifiers described on the previous section, others were tested over the database, as well as a large number of parameters combinations and algorithms combinations (meta-learning ensembles). We show the results of the best algorithms and parameterization. In the next sections, we described the adopted metrics in order to evaluate each built model.

3.5.1 *Kappa.* Kappa metric represents the agreement level between two or more observers that are evaluating the same object, on our work it is the prediction of the pink shrimp harvest. The quantification of the agreement level is measured by Kappa index [Cohen 1960]. Kappa's value are from -1 to 1, on which one represents fully agreement, zero is the expected by chance and values under zero less than expected by chance.

3.5.2 *MinF1 and MaxF1.* F-Score metric [van Rijsbergen 1979] represents the harmonic average between the recall and precision. The MinF1 and MaxF1 are derived from the F-Score [Drews-Jr et al. 2013]. The MaxF1 provides an average on which each class is treated with the same importance, while the MinF1 provides an average on which the instances are treated equally. The MinF1 is more influenced by the performance of the classifier on more populated classes and has the same value of the accuracy, while the MaxF1 is more influenced by classes with less elements. Accuracy and MinF1 are equivalent, thus we adopted accuracy on this work.

3.5.3 *Ordinal Classification Index (OCI).* The present work makes predictions for a class with three possible values: bad, regular and good. These classes have an inherent relation of order (bad $\rightarrow$ regular $\rightarrow$ good) without a numeric difference between them.

The OCI was proposed by [Cardoso and Sousa 2011] with the objective of evaluating a classifier directly from the confusion matrix. The OCI is an error coefficient based on how the obtained result diverges from the ideal prediction. A penalization is applied based on how many elements are out of the main diagonal, the penalization increases according with the distance of the elements from the main diagonal. Smaller values of OCI are expected, while larger values represents bad estimation.

4. EXPERIMENTAL RESULTS

The description of supervised machine learning techniques and algorithms for classification problem, as well as the establishment of the evaluation metrics allow us to proceed to presentation of the results and their analysis. All the models are tested using the cross-validation method [Kohavi 1995b] with ten folds.

Firstly, the use of several algorithms on the experiments involving meta-learning techniques creates the need of a unique terminology for each experiment. This is presented on the Table II.

Table II. Legend for the Meta-Learning Ensembles

Experiment	Meta-Classifier(s)	Classifier(s)
GradingLWL	Grading Bagging (SVN)	DecisionStump, SVN, LWL (DecisionStump)
GradingTree	Grading Bagging (SVN)	SVN REPTree
Voting	Voting	SVN, REPTree LWL (DecisionStump)

Table III presents the results obtained of each experiment using the proposed evaluation metrics. Despite the order of the table is defined by the OCI metric, it is notable the correspondence between all the metrics. Kappa and MaxF1 metric differs only by the position of one classifier (DecisionTable) on the table.

Table III. Metrics results for each classifier (ordered by OCI)

Experiment	Accuracy	Kappa	OCI	MaxF1
GradingLWL	91%	0.81	0.1	0.89
GradingTree	88%	0.73	0.14	0.82
Bagging (SVN)	85%	0.65	0.19	0.77
Voting	83%	0.62	0.22	0.75
DecisionTable	83%	0.65	0.23	0.78
LWL(DecisionStump)	80%	0.62	0.28	0.75
REPTree	76%	0.55	0.3	0.7
JRip	75%	0.54	0.32	0.69

All the classifiers reached at least 75% of accuracy, Figure 1 shows the accuracy for all experiments. The best result was obtained by the GradingLWL ensemble with 91% of accuracy and 0.1 on the OCI.

Accuracy values were higher than their respective MaxF1 values due to the lack of balance between the classes. Accuracy is influenced by the classifier performance on the larger classes. For example, the Bagging (SVN) classifier presented the greatest difference between the Accuracy (0.85) and the MaxF1 (0.77), the confusion matrix is presented on the Table IV. On the largest class (regular with forty elements) only one instance was incorrectly classified (1.66%), while for the other classes four instances were incorrectly classified, representing 6.66% of error for each class.

The GradingLWL is the only method that presented the perfect agreement for Kappa (0.81-0.99), while the classifiers JRip and REPTree presented moderate agreement (0.41-0.60). The other methods were under substantial agreement (0.61-0.80).

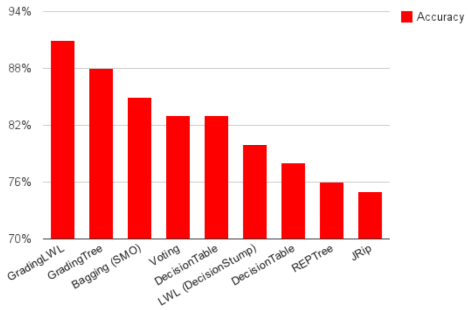

Fig. 1. Accuracy for all experiments

Table IV. Bagging (SVN) Confusion Matrix

		Predicted Class		
		Bad	Regular	Good
Actual Class	Bad	8	4	0
	Regular	1	39	0
	Good	0	4	4

The OCI metric was observed with more careful due to its conception, which fitted perfectly on the analyzed process. Except the Voting method, the other ensemble methods (GradingLWL, GradingTree and Bagging(SVN)) were under the 0.2 margin, with the GradingLWL reaching 0.1. The other four classic classifiers remain in the interval $[0.23; 0.32]$.

Although it was not the objective of the current paper, some attributes were found as relevant on the analyzed process. The algorithms DecisionTable, JRip and REPTree provided meaningful outputs, on which we were able to identify the El Niño anomaly, wind direction and captured shrimps (length $\leq$ 80mm) as directly correlated to the size of the harvest.

5. CONCLUSIONS

This work had as main objective the prediction of the behaviour of the pink shrimp harvest at the Patos Lagoon Estuary (PLE), which has direct impact on the local economy, due to several families having the fishing activity as their main income source. Over the last decades the pink shrimp harvest on the PLE has shown a decreasing tendency year by year and it is of major concern for the researchers at the moment, to keep tracking the harvest volume especially to help on the development of public policies to maintain and protect this species in this estuary life phase.

The data employed to build the models were carefully chosen in order to fulfill the requirement. The first is high reliability, then the availability of the data and the frequency of update. Source of data that have big delay on updates were discarded in favor of sources with less update delay.

The performance of the eight supervised machine learning algorithms was evaluated by a set of metrics. The better four supervised machine learning algorithms were ensembles of classifiers, which were built with meta-learning techniques. The named GradingLWL reached 91% of accuracy and 0.1 on the OCI metric, with less classifiers on the ensemble the GradingTree reached 88% and 0.14 on the same metrics. The composition of meta-learning techniques and classic machine learning algorithms is essential to the quality of the achieved results.

Looking for suitable metrics to evaluate the models could be addressed as future work to obtain a new algorithm or improve existing algorithms. For example, on the optimization phase of the JRip the use of Accuracy or MaxF1 could be an alternative to current adopted rule, on which when a perfect match rule is not found the algorithm returns the most populous class. A similar behavior occurs on the DecisionTable, on which if a perfect match rule is not found, the most frequent class of the table is returned.

REFERENCES

Atkeson, C., Moore, A., and Schaal, S. Locally weighted learning. *AI Review*, 1996.

Box, G. E., Jenkins, G. M., and Reinsel, G. C. *Time series analysis: forecasting and control.* John Wiley & Sons, 2008.

Breiman, L. Bagging predictors. *Machine learning* 24 (2): 123–140, 1996.

Cardoso, J. S. and Sousa, R. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence* 25 (8): 1173–1195, 2011.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1): 37–46, 1960.

Cohen, W. W. Fast effective rule induction. In *ICML*. pp. 115–123, 1995.

D'Incao, F. Pesca e biologia de Penaeus paulensis na Lagoa dos Patos, RS (in portuguese). *Atlantica* 13 (1): 159–169, 1991.

D'Incao, F. and Reis, E. Community-based management and technical advice in Patos Lagoon estuary (Brazil). *Ocean & coastal management* 45 (8): 531–539, 2002.

Diop, H., Keithly, W. R., Kazmierczak, R. F., and Shaw, R. F. Predicting the abundance of white shrimp (litopenaeus setiferus) from environmental parameters and previous life stages. *Fisheries Research* vol. 86, pp. 31–41, 2007.

Drews-Jr, P., Colares, R., Machado, P., Faria, M., Detoni, A., and Tavano, V. Microalgae classification using semi-supervised and active learning based on Gaussian mixture models. *Journal of the Brazilian Computer Society* 19 (4): 411–422, 2013.

Frank, E., Hall, M., and Pfahringer, B. Locally weighted naive bayes. In *UAI*. Morgan Kaufmann, pp. 249–256, 2003.

Garcia, S. P. and Almeida, J. S. Multivariate phase space reconstruction by nearest neighbor embedding with different time delays. *Phys Rev E Stat Nonlin Soft Matter Phys* 72 (027205), 2006.

Garcia, S. P., DeLancey, L. B., Almeida, J. S., and Chapman, R. W. Ecoforecasting in real time for commercial fisheries: the atlantic white shrimp as a case study. *Marine Biology* 152 (1): 15–24, 2007.

Göndör, M. and Breffelean, V. REPTree and M5P for measuring fiscal policy influences on the romanian capital market during 2003–2010. *International Journal of Mathematics and Computers in Stimulation* 6 (4): 378–386, 2012.

Grant, W. E., Matis, J. H., and Miller, W. Forecasting commercial harvest of marine shrimp using a markov chain model. *Ecological Modelling* vol. 43, pp. 183–193, 1988.

Holte, R. C. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11 (1): 63–90, 1993.

Iba, W. and Langley, P. Induction of one-level decision trees. In *ICML*. pp. 235–240, 1992.

Kalekar, P. S. Time series forecasting using holt-winters exponential smoothing. Tech. rep., Kanwal Rekhi School of Information Technology, 2004.

Kittler, J., Hatef, M., Duin, R. P., and Matas, J. On combining classifiers. *IEEE TPAMI* 20 (3): 226–239, 1998.

Kohavi, R. The power of decision tables. In *ECML*. pp. 174–189, 1995a.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*. Vol. 14. pp. 1137–1145, 1995b.

Kotsiantis, S. B. Supervised machine learning: A review of classification techniques. *Informatica* vol. 31, pp. 249–268, 2007.

Kuncheva, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, Inc., 2004.

Paiva, M. P. *Recursos pesqueiros estuarinos e marinhos do Brasil (in portuguese)*. UFC Edições, 1997.

Seewald, A. and Fuernkranz, J. An evaluation of grading classifiers. In *Advances in Intelligent Data Analysis: 4th International Conference*, F. H. et al. (Ed.). Springer, Berlin/Heidelberg/New York/Tokyo, pp. 115–124, 2001.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* 100 (3-4): 441–71, 1987.

Sujjaviriyasup, T. and Pitiruek, K. Agricultural product forecasting using machine learning approach. *Int. Journal of Math. Analysis* 7 (38): 1869–1875, 2013.

van Rijsbergen, C. J. *Information Retrieval*. Butterworth-Heinemann, 1979.

Vapnik, V. *Estimation of Dependences Based on Empirical Data*. Springer, 2006.

Vilalta, R. and Drissi, Y. A perspective view and survey of meta-learning. *Artificial Intelligence Review* vol. 18, pp. 77–95, 2002.