Fifth International Conference
Modelling and Development of Intelligent Systems
June 23 - 25, 2017
"Lucian Blaga" University
Sibiu - Romania

Using Top Down Clustering model for the assessment of anthropogenic pressures and threats on the fish populations

Daniel Hunyadi, Angela Bănăduc, Doru Bănăduc

Abstract

The study is based on the necessity of classification and hierarchization of pressures and threats on fish species, in relation with their intensity and spatial extension, as an essential step for ecological impact assessment and management. The input data of our model are represented by human pressures and threats categories on fish species present in the Lower Danube - Danube Delta - North-western Black Sea convergence area, the spatial frequency and the intensity levels of each categories varying from one (minimum) to five (maximum).

1 Introduction

The identified pressures and threats categories are: anthropogenic geo-morphological changes, anthropogenic hydrological changes, habitat fragmentation and/or loss, aquatic and semi-aquatic vegetation diminishing/disparition, pollution and/or eutrophication, fish pouching and/or overexploitation, alien species introduction, trophic resources diminishing/destruction. The obtained model is a combination of decision tree learning and clustering. We chose as being appropriate the RapidMiner (RM) for implementation of this model.

Assessing the capacity of ecosystems to respond to a huge diversity of natural and/or anthropogenous perturbations or disturbances by resisting structural and functional damage and recovering, the so called resilience [1], it is one of the most attractive intellectual problem for any student in the field of ecology and remain a challenge for all his life after, along the ongoing debates related to how to protect the aquatic ecosystems in terms of their resilience assessment and monitoring. This issue is obviously one which cannot be approached by a single point of view and/or taxa, and/or metric. The time scale is very important in this respect in the context in which the rate at which a natural system returns to a single steady or cyclic state [2] it is time dependent. The working hypothesis in this specific research that the fish fauna structure significant variability and return/recovery can be used as a comparative tool to assess diverse types of aquatic ecosystems resilience in the condition of the human impact presence.

Why Lower Danube River-Danube Delta-North West Black Sea area? This convergence area is a rich, complex and dynamic fish fauna formed of three interdependent subsystems which exhibits a significant level of flexibility and adaptation over geological time [3].

The rest of article is organized as follows. In section 2 we present some literature review regarding to clustering and decision trees. Input data, model definition, model implementation and practical results are presented in section 3. Conclusions and further directions of study can be found in section 4.

2 Literature review

The identified pressures and threats categories are: anthropogenic geo-morphological changes, anthropogenic hydrological changes, habitat fragmentation and/or loss, aquatic and semi-aquatic vegetation diminishing/disparition, pollution and/or eutrophication, fish pouching and/or overexploitation, alien species introduction, trophic resources diminishing/destruction. The obtained model is a combination of decision tree learning and clustering. We chose as being appropriate the RapidMiner (RM) for implementation of this model.

2.1 Clustering

In this section, K-means clustering algorithm and some of the widely known weighted clustering algorithms have been described in brief with their limitations.

K-means [9] is one of the most popular clustering algorithms. K-means is a partitioning method, which creates initial partitioning and then uses iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The algorithm is used to classify a given data set into fixed number of clusters (K). K-means uses the concept of centroid, where a centroid represents the center of a cluster. In the process, K centroids, one for each cluster is defined apriori. Then each object of the data set is assigned a group with the closest centroid. The positions of k centroids are recomputed when all objects have been assigned to any of the clusters. The process is repeated until the centroids are no longer move. One of the limitations of K-means is not differentiating among attributes i.e. each attribute is given same importance in the clustering process. To overcome the limitation, weighted K-means is used with different weights for different attributes.

Al-Harbi et.al. [2] applied simulated annealing to generate weights for weighted K-means. Local optimization problem is one of the problems of simulated annealing techniques and that forms a major limitation in this approach too. Ayan et. al. [3] used information gain as attribute weights but the approach has the inherent drawbacks of the concept of information gain. A genetic cluster algorithm is also proposed by Demiriz et. al. [7] that has difficulty in defining fittest function as required in genetic process. In many real world problems, clustering in stand-alone mode does not provide the desired results.

Semi-supervised clustering [5, 6, 10, 15] is becoming popular with the presence of both labeled and unlabeled data in many practical problems. Semi-supervised clustering uses a small amount of labeled objects (where information about the groups is available) to improve unsupervised clustering algorithms. Existing algorithms for semi-supervised clustering can be broadly categorized into constraint-based and distance-based semi-supervised clustering methods. Constraint-based methods [5, 6, 10] are generally based on pair-wise constraints i.e. pairs of objects labeled as belonging to same or different clusters to facilitate the algorithm towards a more appropriate partitioning of data. In this category, the objective function for evaluating clustering is modified such that the method satisfies constraints during the clustering process. In distance-based approaches [5, 11], an existing clustering algorithm uses a particular distance measure.

Wagstaff et. al. [10] has developed another variant of the k-means algorithm i.e. COP-Kmeans by incorporating background knowledge in the form of instance-level constraints. These instance-level constraints help in identifying which objects should be grouped together. An if-statement is introduced to assign cluster and ensures that none of the constraints are violated when the k-means algorithm groups each object to its closest cluster. However, the major limitation of this algorithm is that it does not allow violation of constraints even if it leads to a more cohesive clustering and leaving them vulnerable to noisy supervision. In order to overcome this limitation, Basu et. al. [6] has proposed pair-wise constraint k-means (PCKmeans) algorithm.

PC-Kmeans algorithm is similar to COP-Kmeans, but the main difference is that this algorithm can violate the constraints with some trade off as penalty for doing so. It tries to come up with a good cluster formation while minimizing the penalty that it incurs. A major limitation of this approach is that it assumes a single metric for all clusters, preventing them from having different

shapes. Bilenko et. al. [5] has proposed metric pair-wise constraint kmeans (MPCK-Means) algorithm to get rid of this limitation. MPCK-Means is considered as one of the most popular semi-supervised clustering algorithms in the recent past. Therefore, the proposed approach has been compared with MPCK-Means in the paper. The proposed approach based on Hyperlink-Induced Topic Search (HITS) algorithm is introduced to overcome the limitations of earlier work.

2.2 Decision trees

Decision tree is a popular classification method that results in a flow - chart like tree structure where each node denotes a test on an attribute value [3] and each branch represents an outcome of the test. Decision tree is a supervised data mining technique. It can be used to partition a large collection of data in to smaller sets by recursively applying two-way and /or multi way splits.

Using the data, the decision tree method generates a tree that consists of nodes that are rules. Each [8] leaf node represents a classification or a decision. The training process that generates the tree is called induction. It has been shown in various studies that employing pruning methods can improve the generalization performance of a decision tree; a loosely stopping criterion is used, letting the decision tree to over fit the training set. Then the over-fitted tree is cut back into a smaller tree by removing sub–branches that are not contributing to the generalization accuracy.

Step 1: Create a node N
Step 2: If samples are all of the same class, C then
Step 3: Return N as a leaf node classify with the class C;
Step 4: If attribute-list is empty then
Step 5: Return N as a leaf node classify with the most common class in samples.
Step 6: Select test-attributes, the attribute among attribute-list with the highest information gain;
Step 7: Label node N with test-attribute;
Step 8: For each known value ai of test-attribute.
Step 9: Grow a branch from node N for the position test attribute = ai;
Step 10: Let Si be the set of samples for which test-attribute = ai;
Step 11: If Si is empty then
Step 12: Attach a leaf classify with the most common class in samples;
Step13: Else attach the node returned by generate-decision tree (Si, attribute-list-attribute);

Each internal node tests an attribute, each branch corresponds to characteristic value, and each leaf node assigns a classification.

3 Model specification

3.1 Input data

The input data of our model are represented by human impact pressures and threats categories presence in Lower Danube (1), Danube Delta (2) and Northwestern Black Sea (3) convergence area, and of the human impact pressures and threats on the fish with quantitative effects, with their intensity level varying from 1 (minimum) to 5 (maximum). The identified pressures and threats categories are: anthropogenic geomorphological changes (A), anthropogenic hydrological changes (B), habitat fragmentation and/or loss (C), aquatic and semiaquatic vegetation diminishing/disparition (D), pollution and/or eutrophication (E), fish pouching and/or overexploitation (F), alien species introduction (G), trophic resources diminishing/disparition (H).

These variables were measured for 115 species. Data were centralized in an EXCEL document.

3.1 Model definition

Our model is built in several steps. In the first step, pre-processing data step, input data are normalized in order to be ready for processing. Then, we use divisive hierarchical clustering in order to obtain the optimum number of clusters.

Normalized input data and the optimum number of clusters are used by k-means algorithm. This algorithm is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. In the last step, we obtained our model which is a combination of decision tree learning and clustering).

3.1 Model implementation and practical results

We chose RapidMiner (RM) for implementation of our model. For the hierarchical clustering we use *TopDownClustering* operator. *KMeans* operator solve the non- hierarchical clustering task offering as output data the clusters. The representation of the clustering solution as a decision tree is realized using *DecisionTree* operator. The implemented processes and the practical results are presented in the next section.

First process is used in order to obtain the optimum number of clusters. It use an import operator named *ReadExcel* which reads an *ExampleSet* from the specified Excel file. The *NominalToNumerical* operator is used for pre-processing data named which changes the type of selected non-numeric attributes to a numeric type.

The *TopDownClustering* operator is used for hierarchical clustering and performs top down clustering by applying the inner flat clustering scheme recursively. Top down clustering is a strategy of hierarchical clustering. The result of this operator is a hierarchical cluster model. The output data for this process is the optimal number of clusters. As we can see in figure 1, for our input data we obtained 19 clusters.

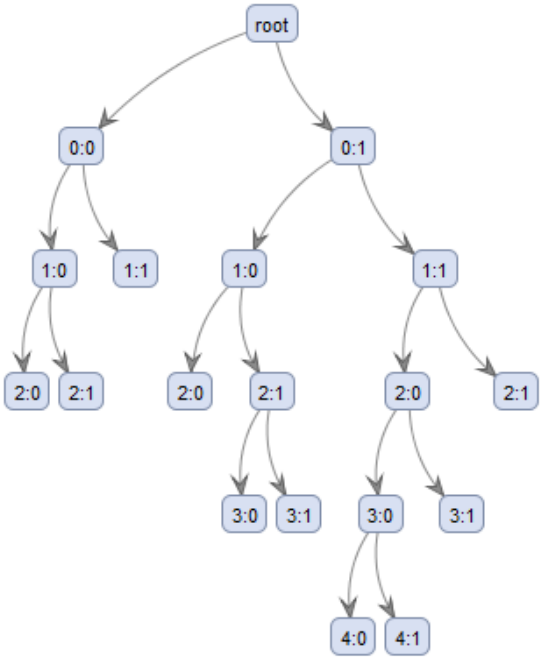

Figure 1. Output data of hierarchical clustering

Second process is a combination of clustering and decision tree and is used to obtain the representation of the model. The chain of the process use *KMeans* operator which performs clustering using the *k-means* algorithm. This operator contains a parameter which specifies the number of clusters to form. The input value for this parameter is the value obtained in the first process.

The *ChangeAttributeRole* operator is used to change the role of one or more attributes. The Role of an attribute reflects the part played by that attribute in an *ExampleSet*. Changing the role of an attribute may change the part played by that attribute in a process. One attribute can have exactly one role. The target role for out attribute is label.

The final operator in our chain is *DecisionTree*. This operator generates a decision tree for classification of both nominal and numerical data. A decision tree is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret.

The chains of the process is presented in figure 2 and the clusters obtained are presented in figure 3.

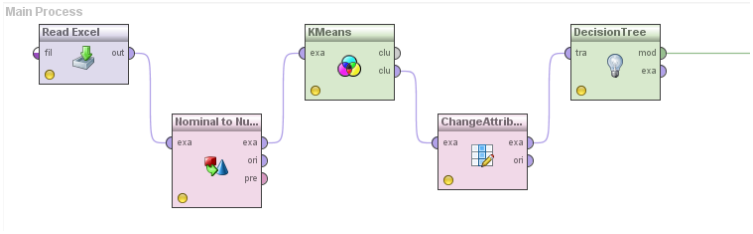

Figure 2. Non-hierarchical process

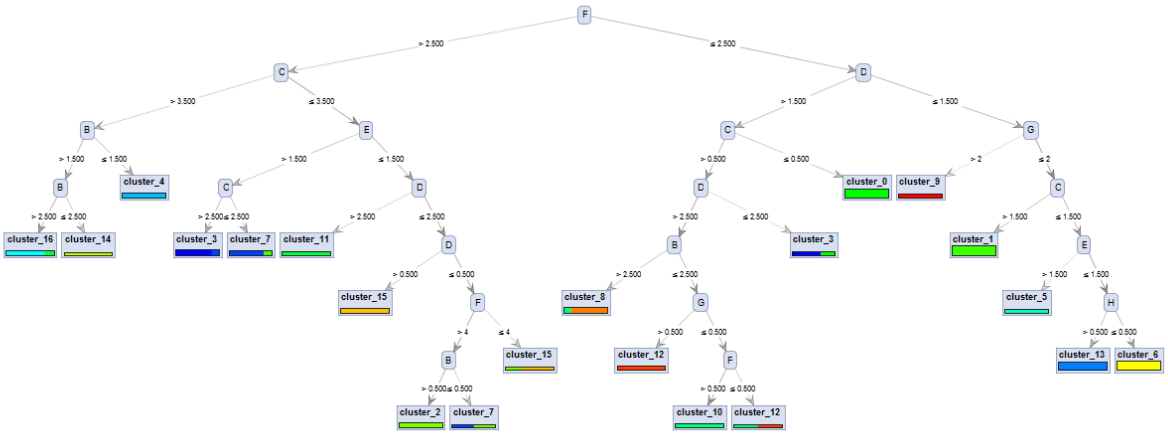

Figure 3. Decision Tree

As we can see in figure 3, the obtained model highlight the fact that the F (fish pouching and/or overexploitation) human impact category has an overall influence on 100% of the studied fish species, in river, delta and marine environments. The studied area was influenced by the human interventions inducing powerful environment alterations, these interventions consisted in large areas embankments for extensive agriculture, intensive fish culture and forestry, with effects on the natural processes, on ecological equilibrium and on the wetlands functions.

Decision trees are powerful and popular for both classification and prediction. The attractiveness of tree-based methods is due largely to the fact that decision trees represent rules. A decision tree model consists of a set of rules for dividing a large heterogeneous population into

smaller, more homogeneous groups with respect to a particular target variable. The intermediate nodes represents the categories for impact pressures. The leaf are the clusters where ale species are segmented. The values from the edges represents the rules for dividing.

4 Conclusions

Based on the 115 fish species and six human impact categories (A - anthropogenic geo-morpholgical changes, B - anthropogenic hydrological changes, C - habitat fragmentation and/or loss, D - aquatic and semi-aquatic vegetation diminishing/disparition, E - pollution and/or eutrophication, F - fish pouching and/or overexploitation, G - alien species introduction, H - trophic resources diminishing/destruction) analysis, the obtained model reveal the connections among some of the human impact categories and variable number of fish.

It is hard to define a mathematical approach to attribute space analysis and description for classification model building because the efficiency of methods and their parameters depend on data structure. Clustering-based decision tree classifier construction ent results and fit certain data structures. Hierarchical clustering is a suitable approach to high density area discovery within classes because of its elasticity choosing cluster defining parameters (similarity measures and linkage options) and it also does not need any prior information about the number of clusters and their positions (centroids). This allows a more objective class structure analysis. The choice of classification model is also relevant because the best model for initial data is not always the best model for data with decomposed classes. This also depends on class structure and the character of overlapping areas. There cannot be a universal approach to cluster combination selection without taking into account the final classification

One possible direction of our study consists in using our proposed methods for other case studies to see their level of generality.

References

[1] Holling, C., 1973. Resilience and stability of ecological systems. Annu Rev Ecol Syst 4:1-23.
[2] Holling, C., 1996. Engineering resilience versus ecological resilience. In Schultze P editor. Engineering within ecological constraints. Washington DC National Academy. 31-44.
[3] Bănăduc D., Planellas S., R., Trichkova T., Curtean-Bănăduc A., 2106. The Lower Danube River-Danube Delta-North West Black Sea: A pivotal area of major interest for the past, present and future of its fish fauna – A short review. Science of The Total Environment, 545-546: 137-151, DOI: 10.1016/j.scitotenv.2015.12.058
[4] S. H. Al-Harbi and V. J. Rayward-Smith, “Adapting k-means for supervised clustering,” Applied Intelligence, vol. 24, pp. 219–226, 2006. [17] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” Journal of the American Statistical Association, 66, 846- 850, 1971.
[5] F. Ayan, “Using information gain as feature weight,” in Proc. of 8th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN’99), Turkey 1999.
[6] S.Balaji and Dr.S.K.Srivatsa” Decision Tree induction based classification for mining Life Insurance Data bases” International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.3, June 2012 pp-699-703.
[7] M. Bilenko, S. Basu and R. Mooney, “Integrating constraints and metric learning in semi-supervised clustering,” in Proc. ICML, 2004, pp. 81–88.
[8] S. Basu, A. Banerjee, and R. J. Mooney, “Active semi-supervision for pairwise constrained clustering,” in Proc. of the 2004 SIAM International Conference on Data Mining (SDM-04), 2004.
[9] A. Demiriz, K. P. Bennett, and M. J. Embrechts, “Semi-supervised clustering using genetic algorithms,” Artificial Neural Networks in Engineering (ANNIE), (1999) 809–814.
[10] Lior Rokach and Oded Maimon” Top-Down Induction of Decision Trees Classifiers- A Survey” IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11, NOVEMBER 2002 pp- 1-12
[11] J. McQueen, Some methods for classification and analysis of multivariate observations, Proc. Symp. Math. Statist. And Probability, 5th, Berkeley, 1 (1967) 281–298.

[12] K. Wagstaff and S. Rogers, "Constrained k-means clustering with background knowledge," in Proc. of 18th International Conference on Machine Learning, 2001, pp. 577–584.
[13] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side information," Advances in Neural Information Processing Systems, vol. 15, pp. 505–512, 2003.

Daniel Hunyadi
"Lucian Blaga" University of Sibiu
Department of Mathematics and Informatics
5-7 Dr. Ratiu Street 550012
Romania
E-mail: daniel.hunyadi@ulbsibiu.ro

Angela Bănăduc
"Lucian Blaga" University of Sibiu
Department of Environmental Sciences, Physics, Physical Education and Sport
5-7 Dr. Ratiu Street 550012
Romania
E-mail: angela.banaduc@ulbsibiu.ro

Doru Bănăduc
"Lucian Blaga" University of Sibiu
Department of Environmental Sciences, Physics, Physical Education and Sport
5-7 Dr. Ratiu Street 550012
Romania
E-mail: doru.banaduc@ulbsibiu.ro